




3 1761 11712364 6



Digitized by the Internet Archive
in 2023 with funding from
University of Toronto

<https://archive.org/details/31761117123646>

12-001

137

Govt

Survey Methodology

Catalogue No. 12-001-XPB

A journal
published by
Statistics Canada

June 2011

•

Volume 37

•

Number 1



Statistics
Canada

Statistique
Canada

Canada

How to obtain more information

For information about this product or the wide range of services and data available from Statistics Canada, visit our website at www.statcan.gc.ca, e-mail us at infostats@statcan.gc.ca, or telephone us, Monday to Friday from 8:30 a.m. to 4:30 p.m., at the following numbers:

Statistics Canada's National Contact Centre

Toll-free telephone (Canada and United States):

Inquiries line	1-800-263-1136
National telecommunications device for the hearing impaired	1-800-363-7629
Fax line	1-877-287-4369

Local or international calls:

Inquiries line	1-613-951-8116
Fax line	1-613-951-0581

Depository Services Program

Inquiries line	1-800-635-7943
Fax line	1-800-565-7757

To access and order this product

This product, Catalogue no. 12-001-X, is available free in electronic format. To obtain a single issue, visit our website at www.statcan.gc.ca and browse by "Key resource" > "Publications."

This product is also available as a standard printed publication at a price of CAN\$30.00 per issue and CAN\$58.00 for a one-year subscription.

The following additional shipping charges apply for delivery outside Canada:

	Single issue	Annual subscription
United States	CAN\$6.00	CAN\$12.00
Other countries	CAN\$10.00	CAN\$20.00

All prices exclude sales taxes.

The printed version of this publication can be ordered as follows:

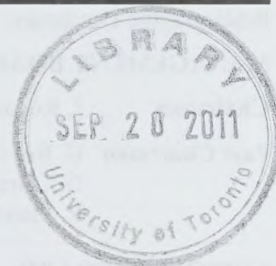
- Telephone (Canada and United States) 1-800-267-6677
- Fax (Canada and United States) 1-877-287-4369
- E-mail infostats@statcan.gc.ca
- Mail
Statistics Canada
Finance
R.H. Coats Bldg., 6th Floor
150 Tunney's Pasture Driveway
Ottawa, Ontario K1A 0T6
- In person from authorized agents and bookstores.

When notifying us of a change in your address, please provide both old and new addresses.

Standards of service to the public

Statistics Canada is committed to serving its clients in a prompt, reliable and courteous manner. To this end, Statistics Canada has developed standards of service that its employees observe. To obtain a copy of these service standards, please contact Statistics Canada toll-free at 1-800-263-1136. The service standards are also published on www.statcan.gc.ca under "About us" > "The agency" > "Providing services to Canadians."

Survey Methodology



A journal
published by
Statistics Canada

June 2011 • Volume 37 • Number 1

Published by authority of the Minister responsible for Statistics Canada

© Minister of Industry, 2011

All rights reserved. This product cannot be reproduced and/or transmitted to any person or organization outside of the licensee's organization. Reasonable rights of use of the content of this product are granted solely for personal, corporate or public policy research, or for educational purposes.

This permission includes the use of the content in analyses and the reporting of results and conclusions, including the citation of limited amounts of supporting data extracted from this product. These materials are solely for non-commercial purposes. In such cases, the source of the data must be acknowledged as follows:

Source (or "Adapted from", if appropriate): Statistics Canada, year of publication, name of product, catalogue number, volume and issue numbers, reference period and page(s). Otherwise, users shall seek prior written permission of Licensing Services, Client Services Division, Statistics Canada, Ottawa, Ontario, Canada K1A 0T6.

June 2011

Catalogue no. 12-001-XPB

Frequency: Semi-annual

ISSN 0714-0045

Ottawa



Statistics
Canada

Statistique
Canada

Canada

SURVEY METHODOLOGY

A Journal Published by Statistics Canada

Survey Methodology is indexed in The ISI Web of knowledge (Web of science), The Survey Statistician, Statistical Theory and Methods Abstracts and SRM Database of Social Research Methodology, Erasmus University and is referenced in the Current Index to Statistics, and Journal Contents in Qualitative Methods. It is also covered by SCOPUS in the Elsevier Bibliographic Databases.

MANAGEMENT BOARD

Chairman J. Kovar

Past Chairmen D. Royce (2006-2009)
G.J. Brackstone (1986-2005)
R. Platek (1975-1986)

Members G. Beaudoin
S. Fortier (Production Manager)
J. Gambino
M.A. Hidioglou
H. Mantel

EDITORIAL BOARD

Editor M.A. Hidioglou, *Statistics Canada*
Deputy Editor H. Mantel, *Statistics Canada*

Past Editor J. Kovar (2006-2009)
M.P. Singh (1975-2005)

Associate Editors

J.-F. Beaumont, *Statistics Canada*
J. van den Brakel, *Statistics Netherlands*
J.M. Brick, *Westat Inc.*
P. Cantwell, *U.S. Bureau of the Census*
R. Chambers, *Centre for Statistical and Survey Methodology*
J.L. Eltinge, *U.S. Bureau of Labor Statistics*
W.A. Fuller, *Iowa State University*
J. Gambino, *Statistics Canada*
B. Hulliger, *University of Applied Sciences Northwestern Switzerland*
D. Judkins, *Westat Inc.*
D. Kasprzyk, *NORC at the University of Chicago*
P. Kott, *National Agricultural Statistics Service*
P. Lahiri, *JPSM, University of Maryland*
P. Lavallée, *Statistics Canada*
P. Lynn, *University of Essex*
D.J. Malec, *National Center for Health Statistics*
G. Nathan, *Hebrew University*
J. Opsomer, *Colorado State University*

D. Pfeiffermann, *Hebrew University*
N.G.N. Prasad, *University of Alberta*
J.N.K. Rao, *Carleton University*
J. Reiter, *Duke University*
L.-P. Rivest, *Université Laval*
N. Schenker, *National Center for Health Statistics*
F.J. Scheuren, *National Opinion Research Center*
P. do N. Silva, *Escola Nacional de Ciências Estatísticas*
P. Smith, *Office for National Statistics*
E. Stasny, *Ohio State University*
D. Steel, *University of Wollongong*
L. Stokes, *Southern Methodist University*
M. Thompson, *University of Waterloo*
V.J. Verma, *Università degli Studi di Siena*
K.M. Wolter, *Iowa State University*
C. Wu, *University of Waterloo*
W. Yung, *Statistics Canada*
A. Zaslavsky, *Harvard University*

Assistant Editors C. Bocci, P. Dick, G. Dubreuil, S. Godbout, D. Haziza, Z. Patak, S. Rubin-Bleuer and
Y. You, *Statistics Canada*

EDITORIAL POLICY

Survey Methodology publishes articles dealing with various aspects of statistical development relevant to a statistical agency, such as design issues in the context of practical constraints, use of different data sources and collection techniques, total survey error, survey evaluation, research in survey methodology, time series analysis, seasonal adjustment, demographic studies, data integration, estimation and data analysis methods, and general survey systems development. The emphasis is placed on the development and evaluation of specific methodologies as applied to data collection or the data themselves. All papers will be refereed. However, the authors retain full responsibility for the contents of their papers and opinions expressed are not necessarily those of the Editorial Board or of Statistics Canada.

Submission of Manuscripts

Survey Methodology is published twice a year. Authors are invited to submit their articles in English or French in electronic form, preferably in Word to the Editor, (smj@statcan.gc.ca, Statistics Canada, 150 Tunney's Pasture Driveway, Ottawa, Ontario, Canada, K1A 0T6). For formatting instructions, please see the guidelines provided in the journal and on the web site (www.statcan.gc.ca).

Subscription Rates

The price of printed versions of *Survey Methodology* (Catalogue No. 12-001-XPB) is CDN \$58 per year. The price excludes Canadian sales taxes. Additional shipping charges apply for delivery outside Canada: United States, CDN \$12 (\$6 × 2 issues); Other Countries, CDN \$20 (\$10 × 2 issues). A reduced price is available to members of the American Statistical Association, the International Association of Survey Statisticians, the American Association for Public Opinion Research, the Statistical Society of Canada and l'Association des statisticiennes et statisticiens du Québec. Electronic versions are available on Statistics Canada's web site: www.statcan.gc.ca.

Survey Methodology
A Journal Published by Statistics Canada
Volume 37, Number 1, June 2011

Contents

Regular Papers

J. Michael Brick, Ismael Flores Cervantes, Sunghee Lee and Greg Norman Nonsampling errors in dual frame telephone surveys.....	1
James O. Chipperfield, Glenys R. Bishop and Paul Campbell Maximum likelihood estimation for contingency tables and logistic regression with incorrectly linked data	13
Yong You and Qian M. Zhou Hierarchical Bayes small area estimation under a spatial model with application to health survey data	25
Hukum Chandra and Ray Chambers Small area estimation under transformation to linearity	39
Sophie Baillargeon and Louis-Paul Rivest The construction of stratified designs in R with the package <i>stratification</i>	53
Jae Kwang Kim and Cindy Long Yu Replication variance estimation under two-phase sampling	67
Stanislav Kolenikov and Gustavo Angeles Cost efficiency of repeated cluster surveys.....	75
Paul Kottnerus On the efficiency of randomized probability proportional to size sampling.....	95

Short Notes

Éric Lesage The use of estimating equations to perform a calibration on complex parameters.....	103
--	-----

In Other Journals	109
--------------------------------	-----

The paper used in this publication meets the minimum requirements of American National Standard for Information Sciences – Permanence of Paper for Printed Library Materials, ANSI Z39.48 - 1984.



Le papier utilisé dans la présente publication répond aux exigences minimales de l'“American National Standard for Information Sciences” – “Permanence of Paper for Printed Library Materials”, ANSI Z39.48 - 1984.



Nonsampling errors in dual frame telephone surveys

J. Michael Brick, Ismael Flores Cervantes, Sunghee Lee and Greg Norman¹

Abstract

Dual frame telephone surveys are becoming common in the U.S. because of the incompleteness of the landline frame as people transition to cell phones. This article examines nonsampling errors in dual frame telephone surveys. Even though nonsampling errors are ignored in much of the dual frame literature, we find that under some conditions substantial biases may arise in dual frame telephone surveys due to these errors. We specifically explore biases due to nonresponse and measurement error in these telephone surveys. To reduce the bias resulting from these errors, we propose dual frame sampling and weighting methods. The compositing factor for combining the estimates from the two frames is shown to play an important role in reducing nonresponse bias.

Key Words: Nonresponse bias; Measurement error; Calibration; Sample allocation; Composite.

1. Introduction

Dual frame telephone surveys that sample from both landline and cell phones have become important in the U.S. to reduce undercoverage bias due to the incompleteness of the landline frame. Blumberg and Luke (2009) show that the percentage of households without a landline telephone but with at least one cell phone has increased dramatically in the last few years, reaching 20 percent by the end of 2008. Other countries also report substantial increases in the percentages of people who have only a cell phone (*e.g.*, Kuusela, Callegaro and Vehovar 2008; Vicente and Reis 2009).

This paper uses data from the California Health Interview Survey (CHIS) and from 8 surveys conducted for the Pew Research Center for the People & the Press to examine the effects of nonsampling errors in dual frame telephone surveys. The CHIS 2007, a survey of California adults, was undertaken in late 2007. It combines a standard landline survey with a screening sample of cell phone numbers, where adults from the cell sample were interviewed only if they indicated that they did not have a landline number in the household. The Pew surveys are national surveys that interviewed an adult at all sampled residential telephone numbers from both landline and the cell samples. These surveys are described in more detail later. A number of important issues associated with the effect of nonsampling errors have been identified as a result of undertaking these dual frame telephone surveys – errors that have not been investigated fully in other studies.

In the next section we review sample design, weighting and variance estimation methods developed for dual frame surveys, and describe CHIS 2007 and Pew dual frame telephone surveys that are used throughout the paper. The

third section discusses nonsampling error in dual frame telephone surveys, and the effects these errors may have on the bias of estimates. Nonresponse and measurement errors have special importance in dual frame surveys. The fourth section studies sampling and estimation methods that may be used to alleviate bias in dual frame telephone surveys, and gives conditions under which these sampling and estimation approaches may be most useful. In this section we propose three estimators to reduce the bias due to differential nonresponse within the overlap domain. The final section summarizes some of the findings for dual frame telephone surveys, and speculates on the applicability of these findings for other dual frame surveys.

2. Background

Most of the literature on dual frame surveys deals with the statistical theory related to efficiency in sample design and estimation. We summarize some of the key results in sampling, weighting and variance estimation, and then discuss the application of these methods to dual frame telephone surveys.

2.1 Sampling

The two sampling frames are denoted as A and B , and we assume the samples from these frames, S_A and S_B , are independent. The domain of units that are only in A is a , the domain of units only in B is b , and the intersection containing the overlap units is ab . In our application to telephone surveys, A is the frame of landline numbers, B is the frame of cell phone numbers, a is the domain of households with only landline numbers, b is the domain of households with only cell phone numbers, and ab is the domain of households with both types of telephone service.

1. J. Michael Brick, Westat and Joint Program in Survey Methodology at the University of Maryland. E-mail: mikebrick@westat.com; Ismael Flores Cervantes, Westat; Greg Norman, Westat, 1600 Research Boulevard, Rockville Maryland, 20850 U.S.A.; Sunghee Lee, Institute for Social Research, University of Michigan, 426 Thompson St. Ann Arbor, MI 48104, U.S.A.

Many important features of dual frame surveys depend on how units that could fall into both sampling frames (ab) are handled.

A screening dual frame approach attempts to make $ab = \emptyset$ by removing any overlap units before sampling, after sampling but prior to data collection, during data collection, or after data collection. Lohr (2009) gives examples of dual frame surveys using each of these approaches.

Brick, Edwards and Lee (2007) and Fleeman (2007) describe screening in dual frame telephone surveys. While U.S. telephone numbers can be partitioned by whether they are cell or landline numbers, this frame does not identify whether those numbers correspond to households with only landlines (a), households with only cell phones (b), or households with both types of service (ab). In the surveys described by Brick, Edwards and Lee (2007) and Fleeman (2007), households sampled from the cell phone frame (B) were screened out during the data collection if they reported having a landline. The CHIS 2007 used this screening approach.

A second approach is called an overlap dual frame survey, and units in the overlap could be sampled from both frames. In this case, estimation methods must be employed to avoid biased estimates because the overlap units have multiple chances of selection. Steeh (2004), Brick, Brick, Dipko, Presser, Tucker and Yuan (2007), and Kennedy (2007) discuss dual frame telephone surveys with overlap. In these cases, all respondents are interviewed irrespective of the frame they are sampled from. The Pew surveys use the overlap approach.

2.2 Estimation

In a screening survey, producing weights for estimating totals and characteristics of the entire population is simple, at least in the absence of nonsampling errors. Since $ab = \emptyset$ and the sampling is independent, the units sampled from each frame are assigned weights that are the inverse of their selection probabilities from the frame from which they were selected. An overall estimate of the total is the sum of the weighted domain estimates, $\hat{y}_{scr} = \hat{y}_a + \hat{y}_b$, where $\hat{y}_a = \sum_{i \in S_a} d_i y_i$ and $\hat{y}_b = \sum_{i \in S_b} d_i \delta_i(b) y_i$, where d_i is the inverse of the selection probability and $\delta_i(b) = 1$ if i is in domain b and 0 otherwise. Variance estimation is also straight-forward since the two frames are strata and variance estimation methods appropriate for stratified samples can be applied. For telephone surveys, the landline sample units are weighted and added to the weighted cell phone sampled units, after the sampled cell phone units that have landlines are given a weight of zero.

Screening during data collection, even in the absence of nonsampling errors, does have implications. For example, screened out households from B are not eligible for the

interview, and this increases data collection costs and the variance of estimated totals (Kish 1965, Chapter 11). The units that are screened out should also be treated properly as sampled units in variance estimation.

Overlap surveys are more complex because units could be sampled from either of the frames. One estimation approach is to combine the two domain estimates, \hat{y}_a and \hat{y}_b with an average of the estimates of the overlap population from the separate frames. If \hat{y}_{ab}^A and \hat{y}_{ab}^B are the weighted estimates of the overlap domain from frame A and frame B , respectively, then an average or composite estimator is $\hat{y}_{ave} = \hat{y}_a + \hat{y}_b + \lambda \hat{y}_{ab}^A + (1 - \lambda) \hat{y}_{ab}^B$, with $0 \leq \lambda \leq 1$. Following Lohr (2009) we refer to these as average estimators. Assuming \hat{y}_a and \hat{y}_b are unbiased for domain a and domain b , and \hat{y}_{ab}^A and \hat{y}_{ab}^B are both unbiased for domain ab , then \hat{y}_{ave} is an unbiased estimator of the total. Estimates of means and other quantities can be produced using weights, where the weights for units in ab that are sampled from A are multiplied by λ and the weights for overlap units sampled from B are multiplied by $(1 - \lambda)$. The choice of the compositing factor, λ , has been investigated by many researchers and specific choices to reduce the variance of the estimates have been suggested by Hartley (1962, 1974) and Fuller and Burmeister (1972). All of average estimators require that the domain for all sampled units can be identified.

Variance estimation with the average estimator is relatively simple if λ is a fixed and not dependent on the selected sample. In this case, $V(\hat{y}_{ave}) = V(\hat{y}_a + \lambda \hat{y}_{ab}^A) + V(\hat{y}_b + (1 - \lambda) \hat{y}_{ab}^B)$, and each of these variances can be computed using variance estimation methods appropriate for the separate samples. If λ is sample dependent, as with the Hartley and Fuller and Burmeister estimators, then variance estimation is more complicated. The average estimators with a fixed λ have been used in most dual frame telephone surveys with overlap. This approach is discussed below for the Pew surveys.

Other estimation approaches that have been considered for an overlap survey include the single frame estimator (Bankier 1986; Kalton and Anderson 1986; and Skinner 1991), and the pseudo-maximum-likelihood estimator (Skinner and Rao 1996; Lohr and Rao 2000; and Lohr and Rao 2006). Lohr (2009) reviews these estimators. Nearly all telephone surveys with overlap that we have seen use some versions of the average estimator, and it is the focus of this research.

2.3 Telephone survey applications

Data from CHIS 2007 are used to illustrate issues that arise in dual frame telephone survey that use a screening approach. The CHIS 2007 is a telephone survey of

California's population conducted by the UCLA Center for Health Policy Research in collaboration with the California Department of Public Health, the California Department of Health Care Services, and the Public Health Institute. Data collection for CHIS 2007 was carried out by Westat in late 2007 through early 2008.

In the CHIS 2007 landline sample, one adult was sampled and interviewed in each household. In the cell phone sample, persons living in households with landline phones were screened out; an adult was sampled and interviewed in the cell sample if they lived in a household classified as cell-only. All responding households, including those screened out from the cell phone frame, were asked questions about telephone status and usage. Nearly 49,000 adult interviews were completed from the landline sample, and 825 interviews were completed with cell-only adults. The landline sample response rate was 35.5% in the interview conducted with a household informant, and a 59.4% for the sampled adult. Respective response rates for the sample from the cell frame were 22.1% and 52.0%. Since CHIS 2007 used a screening approach, the reported response rate for the cell-only household informant interview is 30.5%. California Health Interview Survey (2009) discusses details of the study design, including differences between the overall cell phone response rate and the cell-only rate.

In the CHIS 2007, the estimates from the cell phone sample are calibrated to the cell-only adult population in California at the screening stage (prior to nonresponse weight adjustment for the sampled adult). There are some difficulties with obtaining reliable control totals for the calibration at the state level that are discussed later. The two samples from the two frames are independent samples and are treated as such, until the ultimate stage where the two are combined and calibrated to independent totals of the entire adult population of California. This last calibration stage does not include telephone status as a domain.

For dual frame telephone surveys with overlap, we use data aggregated from 8 surveys conducted for the Pew Research Center for the People & the Press in late 2008 through early 2009. (The data for the Pew surveys were provided by Scott Keeter of the Pew Research Center for the People & the Press). All of these are surveys of the entire U.S. adult population. The surveys interview one adult in each sampled household from both frames using nearly identical questionnaires. Over the 8 surveys, nearly 11,300 landline interviews and 3,800 cell phone interviews were completed. The response rates from the different surveys are very similar for the landline and the cell phone samples, with a median difference of one percentage point between the samples from the two frames. The response rates range across the 8 surveys and two frames from 17% to 24%.

In the Pew surveys, like most dual frame telephone surveys with overlap, a calibrated version of the average estimator is employed. Most surveys calibrate to both the telephone status domain counts (number of adults living in households with only cell phones, the number in household with only landlines, and households with both landlines and cell phones), and to demographic variables. The Pew studies are also calibrated to demographic totals including age, education, race/ethnicity, region, and population density of households with adults 18 years of age or older. In addition, they calibrate to totals of telephone status and, within the overlap domain to relative usage of landline and cell phones.

3. Nonsampling errors

Dual frame theory has been developed for ideal conditions – complete response and the absence of other nonsampling errors. Nonsampling errors affect the bias and precision of the estimates in any survey, but their effects in dual frame surveys may be qualitatively different from those in single frame surveys for three reasons. First, nonsampling error in dual frame surveys often makes it difficult to determine the probability of selection of the sampled unit. This occurs when domain membership is ascertained during data collection, and nonresponse and measurement errors make it difficult to determine if a sampled unit is in the overlap. Second, nonsampling error in dual frame surveys may be linked directly, sometimes causally, to the sampling frame especially when data collection approaches differ by frame. Third, sampling from more than one frame adds complexity and creates more opportunities for nonsampling errors to have differential effects.

3.1 Nonresponse effects

Brick, Dipko, Presser, Tucker and Yuan (2006) show that the over-representation of the number of adults in cell-only households that occurs in almost all dual frame telephone samples may be due to nonresponse error. They suggest that this over-representation might be the result of differential accessibility – adults who rarely use cell phones are less likely to answer their cell phone than those who use their cell phones regularly. They did not find the same type of usage-related differential response rates in the landline sample. Kennedy (2007) further explores this type of nonresponse bias by examining the effects on specific estimates.

To evaluate the differential representation, we compare the CHIS 2007 and Pew survey sample distributions by sampling frame and telephone usage to estimates from the National Health Interview Survey (NHIS). The NHIS is a

face-to-face survey sponsored by the National Center for Health Statistics with data collected by the U.S. Bureau of Census (the NHIS data were provided by S. Blumberg and J. Luke as a special tabulation). It is the only federal government survey that provides estimates of telephone status and usage (Blumberg and Luke 2009). We define usage for the dual users (those in households with both types of phone service) as cell-mainly and land-mainly, where cell-mainly are persons who live in households that receive all or almost all their calls on their cell phone and land-mainly are the dual users in households that do not receive all or almost all their calls on their cell phone.

To be more comparable to the CHIS figures, Table 1 restricts the NHIS estimates to those from the West region only (NHIS estimates for California are not available). California accounts for 52 percent of the adults in the West. The NHIS figures are population estimates from the first six months of 2008, which is roughly contemporaneous to the CHIS data collection period. The CHIS figures are the unweighted sample dispositions (the weighted dispositions are nearly identical). Even though CHIS used a screening

approach, the telephone usage information was collected for every responding household in the cell phone sample. The table shows that the cell phone frame distribution over-represents the percent of adults in cell-only households and under-represents land-mainly adults when compared to the NHIS estimates. The landline respondents over-represent the land-only users and under-represent the cell-mainly dual users. The landline frame differences are more substantial than observed in a 2004 survey as reported in Brick *et al.* (2006).

Table 2 shows the same type of comparison of the NHIS national estimates from the second half of 2008 to the aggregated Pew survey unweighted outcomes (all the surveys were equal probability samples). Similar to the CHIS results, the cell frame distribution from the Pew surveys over-represents the percentage in the cell-only group and under-represents the land-mainly group, but the differences are less substantial than in CHIS. The Pew distribution from the landline sample mirrors the NHIS distribution closely, with a slight under-representation of the cell-mainly group.

Table 1
Percentage distribution of adults from CHIS 2007 and NHIS, by telephone usage

Telephone usage	NHIS West adults in landline households	CHIS 2007 landline distribution	NHIS West adults in cell phone households	CHIS 2007 cell phone distribution
Landline-only	23.5% (1.5%)	34.2% (0.2%)	—	—
Dual – land-mainly	56.6% (1.7%)	53.2% (0.2%)	60.9% (1.7%)	18.5% (0.7%)
Dual – cell-mainly	19.9% (1.4%)	12.7% (0.2%)	21.4% (1.4%)	31.2% (0.9%)
Cell-only	—	—	17.7% (1.3%)	50.3% (0.9%)
Total	100.0%	100.0%	100.0%	100.0%

Notes NHIS-West is the National Health Interview Survey, West Region, first 6 months of 2008, with percentages of all households with that type of service (thanks to S. Blumberg and J. Luke for this special tabulation). CHIS 2007 is the California Health Interview Survey, collected in 2007 and early 2008, with unweighted percentages from the landline and cell frames. In the cell phone sample, usage was obtained in the screening interview. Approximate standard errors given in ().

Table 2
Percentage distribution of adults from Pew surveys and NHIS, by telephone usage

Telephone usage	NHIS adults in landline households	Pew surveys landline distribution	NHIS adults in cell phone households	Pew surveys cell phone distribution
Landline-only	19.4% (0.7%)	23.0% (0.4%)	—	—
Dual – land-mainly	58.8% (0.8%)	62.7% (0.5%)	58.8% (0.8%)	42.3% (0.8%)
Dual – cell-mainly	19.3% (0.7%)	14.4% (0.3%)	18.5% (0.7%)	24.0% (0.7%)
Cell-only	—	—	22.7% (0.7%)	33.7% (0.8%)
Total	100.0%	100.0%	100.0%	100.0%

Notes NHIS is the National Health Interview Survey, second 6 months of 2008, with percentages of all households with that type of service. Pew surveys aggregates 8 surveys conducted for the Pew Research Center for the People & the Press from October 2008 through March 2009, with unweighted percentages from the landline and cell frames. (Thanks to S. Keeter for providing these data). Approximate standard errors given in ().

Both of these surveys exhibit response distributions by frame and usage that are consistent with the accessibility conjecture of Brick *et al.* (2006). This conjecture implies an ordering of those that are most accessible and likely to respond – ordering from the most likely to respond to the least likely to respond in the cell frame is cell-only, cell-mainly, and land-mainly. The special problem due to having two frames is that the ordering in the landline frame is different (land-only, land-mainly, cell-mainly), and the overlap units from the two frames could have very different response rates and biases.

To examine nonresponse bias for a dual frame survey with overlap, suppose both the landline and cell samples are poststratified to telephone status domain totals prior to forming an average overall estimate. The poststratified estimator is

$$\hat{y}_{ps} = \frac{N_a}{\hat{N}_a} \hat{y}_a + \frac{N_b}{\hat{N}_b} \hat{y}_b + \lambda g^A \hat{y}_{ab}^A + (1 - \lambda) g^B \hat{y}_{ab}^B, \quad (1)$$

where the poststratification factor for the land-only sample is N_a / \hat{N}_a , for the cell-only sample it is N_b / \hat{N}_b , and the frame specific poststratification factors for the overlap are $g^A = N_{ab} / \hat{N}_{ab}^A$ and $g^B = N_{ab} / \hat{N}_{ab}^B$ for the landline and cell samples, respectively. The Horvitz-Thompson (HT) estimators of the number of units are \hat{N}_a for the land-only domain, \hat{N}_b for the cell-only domain, and \hat{N}_{ab}^A and \hat{N}_{ab}^B for the overlap domain from the two samples. Since we focus on the overlap, we write

$$\hat{y}_{ps,ab} = \lambda g^A \hat{y}_{ab}^A + (1 - \lambda) g^B \hat{y}_{ab}^B. \quad (2)$$

This poststratified estimator differs from the approach suggested by Lohr and Rao (2000), who average and then poststratify rather than poststratify and then average. Both approaches are consistent and approximately unbiased when there are no nonsampling errors.

If we allow for differential response rates by telephone usage within the overlap such as those observed in dual frame telephone surveys, (2) is biased. Let W be the proportion of the overlap that are land-mainly, and let \bar{Y}_{ml} and \bar{Y}_{mc} be the population means for a characteristic for land-mainly and cell-mainly dual users, respectively. The bias of $\hat{y}_{ps,ab}$ is

$$b(\hat{y}_{ps,ab}) \doteq WN_{ab}(\bar{Y}_{ml} - \bar{Y}_{mc}) \\ (\lambda r_{l1} r_l^{-1} + (1 - \lambda) r_{c1} r_c^{-1} - 1), \quad (3)$$

where r_l is the dual user's response rate for the landline sample, r_{l1} is the landline sample response rate of the land-mainly, r_c is the dual user's response rate for the cell sample, and r_{c1} is the cell phone sample response rate of the land-mainly.

To derive (3), we first define land-mainly and cell-mainly domain estimators from the landline sample as $\hat{y}_{ab}^A(ml) = \hat{N}_{ml}^A \bar{y}_{ab}^A(ml)$ and $\hat{y}_{ab}^A(mc) = \hat{N}_{mc}^A \bar{y}_{ab}^A(mc)$, and from the cell sample as $\hat{y}_{ab}^B(ml) = \hat{N}_{ml}^B \bar{y}_{ab}^B(ml)$ and $\hat{y}_{ab}^B(mc) = \hat{N}_{mc}^B \bar{y}_{ab}^B(mc)$. Now assume (a) $E \bar{y}_{ab}^A(ml) = E \bar{y}_{ab}^B(ml) = \bar{Y}_{ml}$ and $E \bar{y}_{ab}^A(mc) = E \bar{y}_{ab}^B(mc) = \bar{Y}_{mc}$; (b) covariances such as $\text{cov}(\hat{N}_{ml}^A / \hat{N}_{ab}^A, \bar{y}_{ab}^A(ml)) = 0$; and, (c) the expected domain totals are simple expressions such as $E \hat{N}_{ml}^A = r_{l1} N_{ml}$, $E \hat{N}_{ab}^A = r_l N_{ab}$, etc. Since $E(N_{ab} / \hat{N}_{ab}^A) \hat{y}_{ab}^A = N_{ab} E\{(\hat{N}_{ml}^A \bar{y}_{ab}^A(ml) + \hat{N}_{mc}^A \bar{y}_{ab}^A(mc)) / \hat{N}_{ab}^A\}$, we can write $E(N_{ab} / \hat{N}_{ab}^A) \hat{y}_{ab}^A \doteq r_{l1} r_l^{-1} N_{ml} \bar{Y}_{ml} + r_{l2} r_l^{-1} N_{mc} \bar{Y}_{mc} = N_{ab} (r_{l1} r_l^{-1} W(\bar{Y}_{ml} - \bar{Y}_{mc}) + \bar{Y}_{mc})$. A corresponding expression can be written for $E g^B \hat{y}_{ab}^B$. Combining the two gives (3).

These expressions assume that $E \bar{y}_{ab}^B(ml) = \bar{Y}_{ml}$ and $E \bar{y}_{ab}^B(mc) = \bar{Y}_{mc}$. An alternative approach that does not require this assumption is to posit that there is response propensity associated with telephone usage. The bias in this case would be a function of the response propensities from each frame. We do not examine the response propensity approach here.

Expression (3) shows that when $0 < W < 1$, the bias of $\hat{y}_{ps,ab}$ is zero if (a) $\bar{Y}_{ml} = \bar{Y}_{mc}$; or (b) $\lambda r_{l1} r_l^{-1} + (1 - \lambda) r_{c1} r_c^{-1} = 1$. Condition (a) is basically the well-known condition from single frame methodology. Condition (b) differs from single frame expressions because the bias depends on both the relative response rates and the compositing factor, λ . The exception is when $r_{l1} r_l^{-1} = r_{c1} r_c^{-1}$, or equivalently $r_{l1} r_{l2}^{-1} = r_{c1} r_{c2}^{-1}$, where r_{l2} is the landline sample response rate of the cell-mainly and r_{c2} is the cell sample response rate of the cell-mainly. In this form, this expression is comparable to the single frame bias expression that shows no bias exists when response rates are constant.

More generally, the value of λ affects the bias of the estimate, not just its variance. The bias can be eliminated by choosing

$$\lambda_0 = \frac{r_l(r_c - r_{c1})}{r_c r_{l1} - r_l r_{c1}}. \quad (4)$$

Since the proportion of the total population covered by the landline frame is approximately equal to the proportion covered by the cell phone frame, most applications have used $\lambda = 0.50$ without considering its effect on bias.

We can now apply these expressions to evaluate the bias of dual frame telephone estimator for CHIS, assuming the bias is only from differential nonresponse in the overlap. Using the data in Table 1, $W = 0.74$ for the NHIS West region. We approximate $r_{l1} r_l^{-1}$ by the relative poststratification factor that is the ratio of the percentage of the CHIS landline sample classified as land-mainly to the percentage of the NHIS adults in landline households that are land-mainly; $r_{c1} r_c^{-1}$ is computed similarly for the cell phone

quantities. The quantities estimated from CHIS 2007 are given in Table 3, $r_{l1} r_l^{-1} \doteq 1.09$ for the landline sample, and $r_{c1} r_c^{-1} \doteq 0.50$ for the cell sample. As an example, suppose $\bar{Y}_{ml} = 0.3$ and $\bar{Y}_{mc} = 0.5$, then the bias of the estimated percentage based on (3) is approximately 3 percentage points (a relative bias of about 9%) if $\lambda = 0.5$. Using (4), the bias is zero when $\lambda \doteq 0.84$; the bias becomes negative for larger values of λ .

Table 3
Within overlap, relative poststratification factors for CHIS 2007 and Pew surveys

Relative poststratification factors*	CHIS 2007	Pew surveys
$r_{l1} r_l^{-1} \doteq g^A / g_{ml}^A$	1.09	1.07
$r_{l2} r_l^{-1} \doteq g^A / g_{mc}^A$	0.50	0.84
$r_{c1} r_c^{-1} \doteq g^B / g_{ml}^B$	0.74	0.78
$r_{c2} r_c^{-1} \doteq g^B / g_{mc}^B$	2.42	1.51

* Poststratification adjustment factor for telephone usage domain within overlap divided by overlap poststratification factor.

The same computations can be done using the data from the Pew surveys, and the estimates are also shown in Table 3. The parameters differ substantially from those computed from CHIS. Since the Pew studies are national, the NHIS estimate is $W = 0.81$. The ratios of the Pew figures to the NHIS also have lower variability than those from the CHIS, with $r_{l1} r_l^{-1} \doteq 1.07$ and $r_{c1} r_c^{-1} \doteq 0.84$. As a result, the bias is only approximately 1 percentage points when $\lambda = 0.5$. The bias is zero when $\lambda \doteq 0.7$.

To evaluate the biases more completely, estimates of $\bar{Y}_{ml} - \bar{Y}_{mc}$ are needed for characteristics from a dual frame telephone survey rather than making arbitrary assumptions as done in the example above. Blumberg and Luke (2009) give estimates that suggest these differences may be as substantial as the differences between the cell-only and landline population that have been documented extensively elsewhere. However, the NHIS estimates are from a face-to-face survey, not a dual frame telephone survey.

Keeter, Dimock and Christian (2008) give estimated characteristics for dual telephone users by sampling frame, but not in sufficient detail to compute the biases. Keeter's estimates indicate the estimates of dual users from the cell frame might be closer to the NHIS overlap estimates than those from the landline frame. However, since the response rates within the overlap are more variable from the cell frame than from the landline frame, a screening design that aims to reduce bias should exclude dual users from the cell phone frame rather than the landline frame when the cell frame has more variable response rates by frame.

Because of the potential bias in the overlap design, Brick *et al.* (2006) suggest using a screening design that excludes adults in dual usage households if they were sampled from the cell frame. In a screening design, a bias still exists due to the differential nonresponse in the landline sample of dual users by telephone usage. Substituting $\lambda = 1$ into (2) and (3), the bias of $\hat{y}_{scr,ab} = g^A \hat{y}_{ab}^A$ is

$$b(\hat{y}_{scr,ab}) = WN_{ab}(\bar{Y}_{ml} - \bar{Y}_{mc})(r_{l1} r_l^{-1} - 1). \quad (5)$$

The bias for this design and estimator is equivalent to single frame estimators, with the bias vanishing when either $\bar{Y}_{ml} = \bar{Y}_{mc}$ or the landline response rates are the same for the land-mainly and the cell-mainly. Notice that in this design, there is no compositing factor that can be used to control the bias.

The bias of the screener estimator for CHIS 2007 is about half that of the average estimator using $\lambda = 0.50$ (the screener bias is 1.3 percentage points compared to the post-stratified average estimator using $\lambda = 0.50$ with bias of -3.3 points). With the Pew parameters, the bias of the post-stratified average estimator and the screener estimator are nearly equal, with the bias of the screener slightly greater than the poststratified estimator (the screener bias is 1.1 percentage points compared to -0.7 points for the post-stratified overlap).

An issue mentioned earlier is that domain totals for poststratification, even for telephone status alone (land-only, cell-only, and dual domains), are not generally available for state or local area surveys. While small area estimates of the percentage of adults who are cell-only at the state level have been published (Blumberg, Luke, Davidson, Davern, Yu and Soderberg 2009), these do not give small area estimates for all three domains. The situation for telephone usage control totals is even more limited, with only national NHIS estimates published. Since the response rates in the cell frame typically vary by usage, some assumptions about the response rates in the cell sample may be useful to avoid substantial over-representation of cell-only and cell-mainly adults from the cell frame sample when using the overlap design.

3.2 Measurement error effects

In addition to nonresponse, some of the differences in the distributions shown in tables 1 and 2 could be due to measurement error. Before we discuss hypotheses related to measurement error, some of the key procedures in the surveys that could be related to measurement error are discussed. There are fundamental differences in the surveys, such as mode and topic. The NHIS is a face-to-face survey; the CHIS and Pew surveys are telephone surveys. Both NHIS and CHIS are health surveys, while the Pew surveys cover a broad range of topics.

The surveys also use different methods for collecting telephone status and usage. In the NHIS an adult family member is asked to answer questions about telephone status and usage for the entire family in a section of the interview about family characteristics. In the cell phone sample in CHIS 2007, the telephone status items are asked during the household screening, but the usage items are in the sampled adult interview. In the CHIS landline sample and the Pew surveys, the status and usage items are all in one of the last sections of the adult interview. This later placement is possible because no screening is involved.

The sampling of an adult is another procedure that may interact with the measurement process. In the CHIS 2007, an adult is sampled from all adults who share the same cell phone. In the Pew surveys, and most other cell phone surveys, the cell phone is considered a personal device, and the person answering the phone is interviewed. In dual use households, the CHIS and Pew methods may result in different samples of adults.

The greatest potential source of measurement error may be related to differences in the questionnaire items for telephone status and usage in the surveys. The items asked in each survey are given in the appendix. The approaches are quite varied. At least part of the difference in the studies is because the CHIS and Pew surveys are conducted by telephone and have prior information about telephone status.

The items used in all three surveys are derived from items used in a supplement to the Current Population Survey (CPS) in 2004. As discussed in Tucker, Brick and Meekins (2007), cognitive testing and behavioral coding for the supplement identified a number of concerns with the CPS items, especially the usage item. Their testing found that a lack of a specific reference period, not having a code for “half the time,” and difficulty in reporting for other members of the household made the usage item susceptible to measurement error. Tucker *et al.* (2007) also highlight the difficulty respondents had in reporting telephone status and usage for all household members in a single item. In addition, respondents had difficulty with understanding the meaning of “landline,” “regular,” a “working” cell phone, and the difference between using and answering a cell phone.

These issues could affect domain classification, and thus bias estimates. For example, a 23-year-old living with parents might report being cell-only, while the parents might report dual usage. The effects on the estimates of these types of measurement errors in the NHIS and telephone surveys are difficult to predict, but inconsistent reporting in telephone and face-to-face administrations is not unexpected.

Another possible measurement problem is the relationship between reporting telephone usage and the sampling frame from which respondents were selected. The hypothesized

error arises if the respondent, when asked which device they use to receive most of their calls, is more likely to choose the device they are using to do the interview. We do not believe this hypothesis has been tested, but any device effect of this nature would be expected to be in the same direction as the nonresponse effect. A dual user should have a greater likelihood of reporting as cell-mainly if sampled from the cell frame; they should be more likely to report as land-mainly if sampled from the landline. Thus, the bias discussed earlier in the context of nonresponse could be arising due to the combined effect of nonresponse and device effect. Without being able to identify the magnitude of these sources of the bias, methods for reducing bias are unclear.

4. Design and estimation approaches with nonsampling errors

Because of the additional issues at play in dual frame surveys, sampling and estimation methods should be designed to account for the most important sources of error rather than focusing solely on sampling error. In this section we address sample design and estimation choices for dual frame telephone surveys within this larger error structure setting.

4.1 Sample design approaches

A key design decision for a dual frame telephone survey is whether to use a screening or full overlap sample design. We begin by exploring the optimal allocation of the sample for overlap and screening designs appropriate for dual frame telephone surveys when simple random samples are selected independently from the two frames and $N_a > 0$, $N_b > 0$, and $N_{ab} > 0$. We assume throughout that the sample sizes are large enough to ignore the finite population correction factors.

We use a linear expected cost function $E(C) = c_A(n_A + n_B c_B c_A^{-1})$, where c_A is the cost of a landline interview, c_B is the cost of a cell phone interview, and n_A and n_B are the number sampled from frames A and B , respectively. Assuming a constant element variance, σ^2 , the variance of the overlap estimator is $v_{ov}^2 = \sigma^2(N_A(N_a + \lambda^2 N_{ab})n_A^{-1} + N_B(N_b + (1 - \lambda)^2 N_{ab})n_B^{-1})$. The allocation that minimizes the variance with this cost function can be found by standard Lagrangian methods, and is

$$n_{o,A} = E(C) \tau^{-1} \sqrt{c_A^{-1} N_A (N_a + \lambda^2 N_{ab})}$$

$$n_{o,B} = E(C) \tau^{-1} \sqrt{c_B^{-1} N_B (N_b + (1 - \lambda)^2 N_{ab})}, \quad (6)$$

where

$$\tau = \sqrt{c_A N_A (N_a + \lambda^2 N_{ab})} + \sqrt{c_B N_B (N_b + (1 - \lambda)^2 N_{ab})}.$$

For a screening design, a linear cost function appropriate for dual frame telephone surveys is $E(C) = c_A n_A + n_b c_b$, where $c_b = c_B + N_B N_b^{-1} c_s$, n_b is the sampled number of cell-only, and c_s is the cost of screening. The variance of the screening estimator is $v_{sc}^2 = \sigma^2 (N_A^2 n_A^{-1} + N_B N_b n_B^{-1})$. The optimal allocation is just the stratified allocation given by $n_{s,A} = E(C) N_A (c_A N_A + \sqrt{c_A c_b} N_b)^{-1}$ and

$$n_B = \frac{E(C) N_B}{\sqrt{c_A c_b} N_A + c_b N_b},$$

yielding

$$n_h = \frac{E(C) N_b}{\sqrt{c_A c_b} N_A + c_b N_b}$$

cell-only interviews.

With no nonsampling error and a fixed expected cost, the variance for the optimally allocated overlap design is smaller than the variance for the optimally allocated screener design when the cost of screening is large enough so that $\sqrt{c_b} > N_b^{-1} (\tau - N_A \sqrt{c_A})$. When bias is included, the screening design may have smaller mean square error than the overlap design even when this condition holds. In the analysis below, we consider bias but do not account for all the effects of nonsampling error. For example, differential response affects the yield by the sampling frame from which the units are selected thus affecting the allocation and variance of the estimate.

We compare the mean square errors of the screening and overlap designs under the CHIS 2007 parameters given previously. The mean square error is the sum of the variance and the bias squared. The variance is for the overall estimate, but the bias arises only from the overlap under our assumptions. The cost parameters for interviewing and screening cell phones are still not very well-known, but we use ($c_A = 1$, $c_B = 3$, $c_s = 2$) based on information given by Keeter *et al.* (2008) and Edwards, Brick and Grant (2008). The other parameters needed for the comparison are the distribution of the population by telephone status domain, and we approximate national values from the 2008 NHIS national estimates ($N_a = 0.2N$, $N_b = 0.2N$, and $N_{ab} = 0.6N$). In this situation, the variance based on an optimally allocated overlap design with $\lambda = 0.5$ is slightly smaller than the variance for the optimal screening design (the ratio of the variances is 0.976). The variances of the two designs are approximately the same when the cost parameters are such that the screening from frame B is slightly less expensive ($c_A = 1$, $c_B = 3$, $c_s = 1.85$).

The screening approach has smaller mean square error than the overlap design under these conditions because the screening approach reduces the bias of the estimates from -3.3 percentage points to 1.3 points. Even a relatively small bias dominates the mean square error comparison between

the two designs, assuming the bias with the screening approach is half the bias under the overlap design. This is the case because the variances of the overlap and screening designs are so similar. If we instead use the parameters from the Pew surveys, then the mean square error for the overlap design is smaller because its bias is lower than the bias of the screener design.

The allocation to the frames with the overlap approach given by (6) assuming only sampling error is determined by the population parameters, the cost parameters, and the compositing factor. While this is not the optimal allocation when differential response rates are admitted, it is still useful to consider this situation since it is likely to be encountered frequently in practice. In this situation, the bias of $\hat{y}_{ps,ab}$ due to differential nonresponse can be eliminated by choosing λ to satisfy (4). Based on the CHIS parameters, the value that eliminates this bias is $\lambda \doteq 0.84$. If we continue with the cost and population assumptions as above, but set $\lambda = 0.84$, then the optimal allocation given by (6) would select about 75% of the sample from the landline frame. This contrasts with the allocation with $\lambda = 0.5$, in which only 63% is from the landline frame. The choice of the compositing factor is critical. When $\lambda = 0.84$ is used in conjunction with the optimal allocation for the CHIS parameters, the estimator is unbiased and has a variance that is about 5 percent less than the estimator from the optimal screener design.

4.2 Estimation approaches

An approach suggested by Brick *et al.* (2006) is to use a full overlap design with an average estimator for the overlap that is poststratified to telephone usage domain totals, as is done in the Pew surveys. This estimator is unbiased and consistent if the estimates within the domains are unbiased and the domain sample sizes are sufficiently large.

The auxiliary data needed for this poststratification for the entire U.S. are now published regularly from the NHIS. As mentioned above, there are some concerns about using these data as control totals that deserve further study. The control totals needed for this estimator are the number of land-only adults, the number of cell-only adults, and the number of adults who are land-mainly and the number who are cell-mainly (N_{ml} and N_{mc} , respectively). This partitions the dual users into its two components.

An alternative estimator of the overlap total using the same auxiliary data is

$$\begin{aligned} \hat{y}_{sep} = & \frac{N_a}{\hat{N}_a} \hat{y}_a + \frac{N_b}{\hat{N}_b} \hat{y}_b + \lambda_1 g_{ml}^A \hat{y}_{ab}^A(ml) \\ & + (1 - \lambda_1) g_{ml}^B \hat{y}_{ab}^B(ml) \\ & + \lambda_2 g_{mc}^A \hat{y}_{ab}^A(mc) + (1 - \lambda_2) g_{mc}^B \hat{y}_{ab}^B(mc), \end{aligned} \quad (7)$$

where the detailed poststratification factors are $g_{ml}^A = N_{ml} / \hat{N}_{ml}^A$, $g_{mc}^A = N_{mc} / \hat{N}_{mc}^A$, $g_{ml}^B = N_{ml} / \hat{N}_{ml}^B$, $g_{mc}^B = N_{mc} / \hat{N}_{mc}^B$, and $0 \leq \lambda_1 \leq 1$; $0 \leq \lambda_2 \leq 1$. This estimator, like the others considered thus far, is unbiased and consistent in the absence of nonsampling errors. Like (1), the estimates from each frame are poststratified before being averaged. The primary difference between (1) and (7) is that the dual users in (7) are partitioned and poststratified by usage; it also introduces different compositing factors within the overlap.

The estimator \hat{y}_{sep} may be useful when (1) is biased and usage control totals are available for poststratification. If the expected means within the usage domains are approximately equal ($E\bar{y}_{ab}^A(ml) = E\bar{y}_{ab}^B(ml) = \bar{Y}_{ml}$ and $E\bar{y}_{ab}^A(mc) = E\bar{y}_{ab}^B(mc) = \bar{Y}_{mc}$), then (7) is unbiased for any choice of $0 \leq \lambda_1 \leq 1$ and $0 \leq \lambda_2 \leq 1$. Since bias is not affected by the choice, different compositing factors may be used to reduce the variance of the estimates as is traditionally suggested in the dual frame literature. Table 3 shows that the proportion of respondents in the detailed usage domains varies considerably by the sampling frame, and this might make different compositing factors worthwhile.

Because telephone usage control totals often are not available, we explored modifying (2) to use different compositing factors similar to those used in the overlap for (7). In this case, the goal would be to reduce bias rather than variance. A modified estimator of the overlap total is

$$\hat{y}_{mod,ab} = \lambda_1 g^A \hat{y}_{ab}^A(ml) + (1 - \lambda_1) g^B \hat{y}_{ab}^B(ml) + \lambda_2 g^A \hat{y}_{ab}^A(mc) + (1 - \lambda_2) g^B \hat{y}_{ab}^B(mc). \quad (8)$$

However, this estimator may not be useful for reducing bias. Earlier, we showed that the bias of $\hat{y}_{ps,ab}$ vanishes when $\lambda_0 = r_l(r_c - r_{cl})(r_c r_{l1} - r_l r_{cl})^{-1}$. The choice of $\lambda_1 = \lambda_2 = \lambda_0$ in (8) eliminates the bias for both land-mainly and cell-mainly estimates, so that different compositing factors are not useful for bias reduction. The bias of the modified estimator is

$$b(\hat{y}_{mod,ab}) = WN_{ab}(\bar{Y}_{ml}(\lambda_1 r_{l1} r_l^{-1} + (1 - \lambda_1) r_{cl} r_c^{-1} - 1) - \bar{Y}_{mc}(\lambda_2 r_{l1} r_l^{-1} + (1 - \lambda_2) r_{cl} r_c^{-1} - 1)), \quad (9)$$

where we make assumptions similar to those used earlier to approximate the bias of $\hat{y}_{ps,ab}$.

Another reason for studying an overlap estimator like (8) is because it is appropriate with sample designs that screen out land-mainly adults from the cell frame. This approach has been considered because the number of cell frame respondents that are classified as land-mainly may be small, and the assumption that $E\hat{y}_{ab}^B(ml) = \bar{Y}_{ml}$ may not hold and biases might result.

Setting $\lambda_1 = 1$, (8) reduces to

$$\hat{y}_{mod\lambda=1,ab} = g^A \hat{y}_{ab}^A(ml) + \lambda_2 g^A \hat{y}_{ab}^A(mc) + (1 - \lambda_2) g^B \hat{y}_{ab}^B(mc). \quad (10)$$

In this design, the landline sample alone is used to estimate both the land-only and the land-mainly totals. Both frames are used to estimate totals for the cell-mainly. If we assume $E\bar{y}_{ab}^A(ml) = \bar{Y}_{ml}$ and $E\bar{y}_{ab}^A(mc) = E\bar{y}_{ab}^B(mc) = \bar{Y}_{mc}$, then we no longer need $E\bar{y}_{ab}^B(ml) = \bar{Y}_{ml}$ for (10) to be unbiased. As before, setting $\lambda_2 = r_l(r_c - r_{cl})(r_c r_{l1} - r_l r_{cl})^{-1}$ eliminates the bias in the cell-mainly estimate.

5. Discussion

This exploration of nonresponse and measurement errors in dual frame telephone surveys suggests the effects of these errors may be very important. It leads us to believe that research on nonsampling errors to reduce biases may be more important than research that leads to incremental reductions in sampling error.

The research also reveals shortcomings in our knowledge about nonsampling errors in these surveys. The direction and magnitude of the effects of measurement error are especially unclear. The inconsistencies in some of the findings for the CHIS 2007 and Pew surveys may well be due to measurement errors associated with the different approaches to data collection in these surveys, or to interactions due to the procedures. A thorough investigation of the error sources in dual frame telephone surveys is essential to improve the quality of dual frame telephone surveys, and we believe experiments to assess the effects of measurement error would be especially beneficial.

We did find that the CHIS 2007 and Pew surveys consistently over-represented cell-only and cell-mainly users in samples from the cell phone frame, and the surveys had a slight over-representation of the land-only and land-mainly from the landline frame. However, the degree of over-representation of the domains differed by survey. In the CHIS, the over-representation could have led to substantial biases in the estimates if an overlap survey and a simple average estimator were used. The CHIS used a screening approach to reduce this potential bias, and this appears to have been largely successful. In the Pew surveys, the representation was less differential by frame and the potential for bias was smaller. In these conditions, the overlap approach may have smaller mean square error than a screening approach.

Due to the potential for bias in dual frame telephone surveys with response patterns like the CHIS 2007, we examined sampling and estimation methods that could be implemented to deal with these biases. We found that screening approaches may be competitive or even preferable in dual frame telephone surveys when the bias due

to differential nonresponse or measurement error is large. If the bias is not negligible, this finding even holds with small sample sizes. However, these results depend on the choice of the compositing factor and the current practice of choosing $\lambda = 0.5$ should be reconsidered. An alternative is to choose the compositing factor to eliminate the bias of the average estimator. In many cases, this approach not only eliminates the bias, but also may be more efficient.

We examined three estimators that deal with the bias due to differential nonresponse within the overlap domain. The first is \hat{y}_{ps} , which uses telephone status as domain control totals. This estimator eliminates the bias due to differential nonresponse when λ_0 is used as the compositing estimator. This compositing factor indirectly uses information on the land-mainly and cell-mainly domain totals in computing response rates by domain and frame. A second estimator, \hat{y}_{sep} , eliminates this source of bias more directly by poststratifying to telephone status and usage control totals. This estimator also permits the use of different compositing factors within the overlap domain to reduce the variance of the estimates. The third estimator that might be used to reduce bias is \hat{y}_{mod} , but this estimator is more pertinent for a sample design that interviews the cell-only and the cell-mainly respondents from the cell frame, along with all respondents from the landline sample. This modified screening design and estimator might be especially attractive if there is concern that the mean of the land-mainly respondents from the cell frame sample is subject to nonresponse bias. All of these estimators could also be raked to additional demographic control totals after combining the two samples.

Given our current state of knowledge, we believe there are important advantages with the full overlap design and \hat{y}_{ps} with λ_0 chosen based on other similar surveys. It is worth observing that even though the CHIS and Pew surveys had very different response patterns, choosing a value of $\lambda_0 = 0.75$ would have reduced the bias substantially for both surveys. An advantage of this estimator over \hat{y}_{sep} in general is that \hat{y}_{ps} is not poststratified to usage domain totals. We suspect that usage domain totals estimated from a face-to-face survey (NHIS) may be subject to substantially different errors than the estimates from telephone surveys. These differences could result in telephone survey estimates that are biased and have underestimated variances. For state and local surveys where even telephone status totals are not well-known, control totals for usage domains are likely to be highly suspect.

A screening design with \hat{y}_{scr} as the estimator has the advantage that it only requires control totals for the entire population and for the cell-only component, such as those estimated from the NHIS. A disadvantage is that, unlike the overlap estimators, there is no compositing parameter that

can be used to reduce the bias directly. The more elaborate screening design that interviews cell-only and cell-mainly from the cell frame and uses \hat{y}_{mod} has merit, but there have been no studies that examine the conditions which would favor this estimator.

A more complete analysis of the effects of nonsampling error would include other factors such as the effect of the differential response rates by frame. For example, we noted that samples from the cell phone frame yield more cell-only households than would be expected. These differential response rates can be addressed in allocating the sample, but we have not done so here. Our exploration of this shows that it results in larger allocations to the landline frame, increases the value of the compositing factor, and makes the screening designs more efficient relative to the overlap designs. The screening design and estimator are still subject to the bias noted above.

While this research concentrated on nonsampling errors in dual frame telephone surveys, we suspect that similar issues exist in many other dual frame surveys, but that these issues may not be recognized. Lohr (2009) mentions nonsampling errors in general dual frame surveys and suggests comparing estimates of the overlap from each frame as a simple diagnostic test. We believe this is an excellent way to begin an investigation of problems associated with the overlap.

As we noted earlier, the handling of the overlap is a major concern in dual frame surveys because nonsampling error may be associated with the sampling frame. Our investigation shows that nonresponse and measurement errors are tied to the sampling frame in dual frame telephone surveys. It is very likely that dual frame telephone surveys that use different modes might experience analogous effects. For example, consider a dual frame household survey designed to survey members of a rare population. Suppose it uses an incomplete membership list with telephone numbers for the rare group as frame *A*, and an area probability sample of households as frame *B*. Different response rates by sampling frame within the overlap might be expected, and these might be related to characteristics of the respondents leading to biases. Even within the overlap, there may be differences such as those related to how long the person has been a member of the organization used to create frame *A* and this might be related to characteristics such as age. This type of situation might parallel some of the within overlap domain issues identified in telephone surveys. Differential measurement errors related to the modes are also possible.

Given the potential for bias in a dual frame survey, one of the important findings of our research is that the compositing factor, λ , influences the bias as well as having an effect on the variance. While the choice of λ typically has only a slight effect on the variance if λ is in the vicinity of

the optimal value, the bias may be more sensitive to this choice. Thus, in dual frame surveys understanding how the choice of λ affects the bias and the mean square error of the estimates is an important consideration. The other sampling and estimation methods discussed in this paper may also be applicable to other dual frame surveys. The usefulness of these methods depends upon understanding the nature of the nonsampling errors as well as the availability of auxiliary data that could be used in calibration.

Acknowledgements

We would like to thank Scott Keeter, Stephen Blumberg and Julian Luke for providing data for this paper. We would also like to thank many people for helpful comments on earlier drafts including Sherm Edwards, Ralph DiGaetano, David Grant, David Hubble, Paul Lavrakas, Graham Kalton, Scott Keeter, and Courtney Kennedy.

Appendix

Telephone usage items

National Health Interview Survey

- N1. *Is there at least one telephone inside your home that is currently working and is not a cellular phone?*
- N2. *Does anyone in your family have a working cellular telephone?*
- N3. *How many working cellular telephones do people in your family have?*
[If both N1 and N2 are 'yes' ask N4]
- N4. *Of all the telephone calls that your family receives, are ...*
All or almost all calls received on cell phones?
Some received on cell phones and some on regular phones?
Very few or none received on cell phones?

California Health Interview Survey – Cell phone

- CC1. *Is this cell phone your only phone or do you also have a regular telephone at home?*
[If the phone is a cell phone and they have a regular phone then ask CC2]
- CC2. *Of all the telephone calls that you receive, are ...*
All or almost all calls received on cell phones
Some received on cell phones and some on regular phones, or

Very few or none on cell phones?

[If respondent replies about half, record it]

California Health Interview Survey – Landline

- CL1. *Do you have a working cell phone?*
[If yes or they share a cell phone ask CL2]
- CL2. *Of all the telephone calls that you receive, are ...*
All or almost all calls received on cell phones
Some received on cell phones and some on regular phones, or
Very few or none on cell phones?
[If respondent replies about half, record it]

Pew Research Center for the People & The Press – Cell phone

- PC1. *Now thinking about your telephone use... Is there at least one telephone INSIDE your home that is currently working and is not a cell phone?*
[If yes ask PC2]
- PC2. *Of all the telephone calls that you receive, do you get?*
[Rotate options—keeping SOME in the middle]
All or almost all calls on a cell phone
Some on a cell phone and some on a regular home phone
All or almost all calls on a regular home phone

Pew Research Center for the People & The Press – Landline

- PL1. *Now thinking about your telephone use... Do you have a working cell phone?*
[If yes ask PL2]
- PL2. *Of all the telephone calls that you receive, do you get?*
[Rotate options—keeping SOME in the middle]
All or almost all calls on a cell phone
Some on a cell phone and some on a regular home phone
All or almost all calls on a regular home phone

References

- Bankier, M.D. (1986). Estimators based on several stratified samples with applications to multiple frame surveys. *Journal of the American Statistical Association*, 81, 1074-1079.

- Blumberg, S.J., and Luke, J.V. (2009). Wireless Substitution: Early Release of Estimates from the National Health Interview Survey, July-December 2008. National Center for Health Statistics. Available at <http://www.cdc.gov/nchs/nhis.htm>.
- Blumberg, S.J., Luke, J.V., Davidson, G., Davern, M.E., Yu, T. and Soderberg, K. (2009). Wireless substitution: State-level estimates from the National Health Interview Survey, January-December 2007. Hyattsville, MD: National Center for Health Statistics. *National Health Statistics Reports*, 14.
- Brick, J.M., Brick, P.D., Dipko, S., Presser, S., Tucker, C. and Yuan, Y. (2007). Cell phone survey feasibility in the U.S.: Sampling and calling cell numbers versus landline numbers. *Public Opinion Quarterly*, 71, 29-33.
- Brick, J.M., Dipko, S., Presser, S., Tucker, C. and Yuan, Y. (2006). Nonresponse bias in a dual frame sample of cell and landline numbers. *Public Opinion Quarterly*, 70, 780-793.
- Brick, J.M., Edwards, W.S. and Lee, S. (2007). Sampling telephone numbers and adults, interview length, and weighting in the California Health Interview Survey cell phone pilot study. *Public Opinion Quarterly*, 71, 793-813.
- California Health Interview Survey (2009). CHIS 2007 Methodology Series: Report 4 – Response Rates. Los Angeles, CA: UCLA Center for Health Policy Research. Available at www.chis.ucla.edu/pdf/CHIS2007_method4.pdf.
- Edwards, W.S., Brick, J.M. and Grant, D. (2008). Relative Costs of a Multi-frame, Multi-mode Enhancement to an RDD Survey. Presented at the Annual Conference of the American Association for Public Opinion Research, New Orleans, LA.
- Fleeman, A. (2007). Survey Research Using Cell Phone Sample: Important Operational and Methodological Considerations. Presented at the Annual Conference of the American Association for Public Opinion Research, Anaheim, CA.
- Fuller, W.A., and Burmeister, L.F. (1972). Estimators for samples selected from two overlapping frames. *Proceedings of the Social Statistics Section*, 245-249.
- Hartley, H.O. (1962). Multiple Frame Surveys. *ASA Proceedings of the Social Statistics Section*, 203-206.
- Hartley, H.O. (1974). Multiple frame methodology and selected applications. *Sankhyā*, C, 36, 99-118.
- Kalton, G., and Anderson, D.W. (1986). Sampling Rare Populations. *Journal of the Royal Statistical Society*, A 149, 65-82.
- Kish, L. (1965). *Survey Sampling*. New York : John Wiley & Sons, Inc.
- Kennedy, C. (2007). Evaluating the effects of screening for telephone service in dual frame RDD Surveys. *Public Opinion Quarterly*, 71, 750-771.
- Keeter, S., Dimock, M. and Christian, L. (2008). Calling Cell Phones in '08 Pre-election Polls. News Release from The Pew Research Center for the People & the Press. Available at www.pewresearch.org/pubs/1061/cell-phones-election-polling.
- Kuusela, V., Callegaro, M. and Vehovar, V. (2008) The influence of mobile telephones on telephone surveys. In *Advances in Telephone Survey Methodology*, (Eds., J.M. Lepkowski, C. Tucker, J.M. Brick, E.D. de Leeuw, L. Japac, P.J. Lavrakas, M.W. Link and R.L. Sangster), New York: John Wiley & Sons, Inc., Chapter 4, 87-112.
- Lohr, S. (2009). Multiple frame surveys. In *Handbook of Statistics: Sample Surveys Design, Methods and Applications*, (Ed., D. Pfeffermann). Elsevier, Amsterdam, Chapter 4, Vol. 29A.
- Lohr, S., and Rao, J.N.K. (2000). Inference in dual frame surveys. *Journal of the American Statistical Association*, 95, 271-280.
- Lohr, S., and Rao, J.N.K. (2006). Estimation in multiple-frame surveys. *Journal of the American Statistical Association*, 101, 1019-1030.
- Skinner, C.J. (1991). On the efficiency of raking ratio estimation for multiple frame surveys. *Journal of the American Statistical Association*, 86, 779-784.
- Skinner, C.J., and Rao, J.N.K. (1996). Estimation in dual frame surveys with complex designs. *Journal of the American Statistical Association*, 91, 349-356.
- Steeh, C. (2004). A New Era for Telephone Surveys. Presented at the Annual Conference of the American Association for Public Opinion Research, Phoenix, AZ.
- Tucker, C., Brick, J.M. and Meekins, B. (2007). Household telephone service and usage patterns in the U.S. in 2004: Implications for telephone samples. *Public Opinion Quarterly*, 71, 3-22.
- Vicente, P., and Reis, E. (2009). The mobile-only population in Portugal and its impact in a dual frame telephone survey. *Survey Research Methods*, 3, 105-111.

Maximum likelihood estimation for contingency tables and logistic regression with incorrectly linked data

James O. Chipperfield, Glenys R. Bishop and Paul Campbell¹

Abstract

Data linkage is the act of bringing together records that are believed to belong to the same unit (*e.g.*, person or business) from two or more files. It is a very common way to enhance dimensions such as time and breadth or depth of detail. Data linkage is often not an error-free process and can lead to linking a pair of records that do not belong to the same unit. There is an explosion of record linkage applications, yet there has been little work on assuring the quality of analyses using such linked files. Naively treating such a linked file as if it were linked without errors will, in general, lead to biased estimates. This paper develops a maximum likelihood estimator for contingency tables and logistic regression with incorrectly linked records. The estimation technique is simple and is implemented using the well-known EM algorithm. A well known method of linking records in the present context is probabilistic data linking. The paper demonstrates the effectiveness of the proposed estimators in an empirical study which uses probabilistic data linkage.

Key Words: Data linkage; Probabilistic linkage; Maximum likelihood; Contingency tables; Logistic regression.

1. Introduction

Data linking, also referred to as data linkage or record linkage, is the act of bringing together records that are believed to belong to the same unit (*e.g.*, a person or business), from two or more files. Data linkage is an appropriate technique when data sets must be joined to enhance dimensions such as time and breadth or depth of detail. Ideally, the linkage will be perfect, meaning only records belonging to the same unit are linked and all such links are made. However, in many situations this does not happen, especially when linking records using fields that may have incorrect values, missing values or values that are legitimately different for a given unit.

Probabilistic linking is often used when the files contain a set of common variables or fields that constitute partial identifying information, but which do not constitute a unique unit identifier. In probabilistic linking (Fellegi and Sunter 1969) all possible links are given a score based on the probability that the records belong to the same unit. This score is calculated by comparing the values of linking variables that are common to both files. A link is then declared if the link score is higher than some cut-off. An optimisation algorithm may be used to ensure that each record on one file is linked to no more than one record on the other file. Probabilistic methods for linking files are now well established (see Herzog, Scheuren and Winkler 2007, Winkler 2001 and Winkler 2005) and there is a range of computer packages available to implement them.

This is a consequence of the continued importance of linkage in a variety of fields, particularly relating to health and social policy. Recent examples of probabilistic data

linkage from the Australian Bureau of Statistics (ABS) include linking records from the 2006 Australian Census of Population and Housing to a number of data sets including Australian death registrations (Australian Bureau of Statistics 2008), the 2006 Census Dress Rehearsal (Solon and Bishop 2009), and the Australian Migrants Settlements Database (Wright, Bishop and Ayre 2009). In the health arena within Australia, probabilistic linkage methods are used by the Western Australian Data Linkage Unit (Holman, Bass, Rouse and Hobbs 1999) and by the New South Wales Centre for Health Record Linkage. Internationally, probabilistic methods are used by Statistics Canada (Fair 2004), USBC (see Winkler 2001), the U.S. National Center for Health Statistics (National Center for Health Statistics 2009) and by the Switzerland Statistical agency as part of their Longitudinal Study of People Living in Switzerland.

Data linking offers opportunities for new statistical output and analysis. Naively treating a probabilistically-linked file as if it was perfectly linked will, in general, lead to biased estimates. Lahiri and Larsen (2005) and Scheuren and Winkler (1993) proposed methods to calculate unbiased estimates of coefficients for a linear regression model under probabilistic record linkage. More recently, Chambers, Chipperfield, Davis and Kovačević (2009) and Chambers (2008) extended this work to a wide set of models using generalised estimating equations and, in the case of linking two files, allowing one file to be a subset of the other file.

This paper develops a maximum likelihood (ML) approach for analysis of probabilistically-linked records. The estimation technique is simple and is implemented using the well-known EM algorithm. The approach involves replacing the statistics, which would be observed from perfectly linked

1. James O. Chipperfield, Australian Bureau of Statistics. E-mail: james.chipperfield@abs.gov.au; Glenys R. Bishop, The Australian National University; Paul Campbell, Australian Bureau of Statistics.

data, with their expectation conditional on the linked data. Assuming this expectation is correctly specified, this approach overcomes the following two limitations of the previous work.

First, the previous methods assume only one linkage pass is made, whereas, probabilistic linkage usually involves multiple passes. In the latter case, records not linked in the first pass are eligible to be linked in the second pass, and only records not linked in the first two passes are eligible to be linked in the third pass, and so on. Each pass is designed to link records with a particular common set of characteristics. For example, the first pass may be designed to link records belonging to individuals who have not changed address between the reference dates of the two files. The second pass may be designed to accommodate changes of address. An example of such an approach is given in Table 1 in section 5.

Second, the previous methods assume that either the two files contain records from exactly the same units or the set of units on one file is a subset of those on the other file. The approach proposed can be used when one of the files to be linked is not necessarily a subset of the other file. This situation occurs frequently in practice and occurred in all the ABS examples mentioned above. It is also worth mentioning that the files to be linked do not need to be related via a sampling mechanism, such as the smaller file being a random sub-sample of individuals from the larger file. Removing this restriction means that the two files may be administrative data sets.

Consider linking two files denoted by X and Y . File Y contains the variable y on the population of individuals U_y comprising n_y records. File X contains a vector of variables, \mathbf{x} , on the population of individuals U_x comprising n_x records. The target of inference is with respect to the population of n_{xy} individuals, denoted by $U_{xy} = U_x \cap U_y$, who are common to File X and File Y . Files X and Y also contain a vector of fields, denoted by \mathbf{z} , which are used to link the files using a probabilistic linkage algorithm. Of course, since we are considering probabilistic linkage here, the variable \mathbf{z} does not constitute a unique unit identifier.

Linking Files X and Y allows the joint distribution of \mathbf{x} and y to be analysed. There are two sources of error that may affect analysis of the joint distribution using the linked file. These errors are referred to as *incorrect links* and *unlinked records*.

A link is correct when the pair of linked records belong to the same individual. A link is incorrect when a pair of linked records do not belong to the same individual. Incorrect links can artificially increase or decrease the correlation between \mathbf{x} and y . An example of the latter is random linkage, where records on File X are randomly linked to records on File Y .

The i^{th} record on File X is defined as an *unlinked record*, if $i \in U_{xy}$ and record i was not linked to a record on File Y . Or in other words, an unlinked record is a record on File X that could be correctly linked but was not linked at all (throughout this paper we use the convention of defining unlinked records in terms of File X , though the definition could equally be in terms of records on File Y). It may not always be possible to link a particular record on File X with much confidence that the link is correct. This situation may arise if a record is missing fields that are useful in establishing the correct link. More generally, unlinked records may occur when some sub-populations are relatively difficult to link. For example, fields such as marital status, qualification, field of study, and highest level of schooling would generally not be as powerful when linking children as when linking mature adults. In this situation, the data linker must decide whether or not to link such records. We define the set of linked records by U_l of size n^* so that $n^* \leq n_x$ and $n^* \leq n_y$.

The problem of analysis with unlinked records has clear parallels with the problem of unit non-response. Both lead to only a subset of legitimate records being available for analysis. The non-response mechanism in survey sampling is, in reality, a function of an unknown set of variables. Here however, we have the slight advantage in knowing that the probability of a record remaining unlinked can only be a function of \mathbf{z} . The problem of non-response is often addressed by weighting or by some conditioning argument. This paper considers both approaches to address the issue of unlinked records.

There is a natural trade-off between the number of unlinked records and incorrect links (and consequently the bias that they introduce). Consider the case where File X is a subsample of File Y so that $U_{xy} = U_x$. Linking all records on File X will result, by definition, in no unlinked records but will result in the number of incorrect links being maximised. If instead we decide to only form links which we are very confident are correct, the number of incorrect links will decrease but the number of unlinked records will increase. In practice, finding the optimal balance between the biases due to unlinked records and incorrect links depends upon the analysis to be undertaken, the linkage methodology, and their interaction. For an in-depth practical discussion of this issue see Bishop (2009).

It is worthwhile mentioning that the problem of making inference in the presence of incorrect record linkage is similar to the problem of making inference in the presence of misclassification of the outcome variable, which is a form of measurement error (see Fuller 1987). In the latter case, identifying assumptions separate the misclassification mechanism from the model mechanism and are required since no error-free measurement is typically available. For example,

Hausman, Abrevaya and Scott-Morton (1998) considers misclassification in the outcome variable of a logistic regression model. Their identifying assumption is that the value of the, possibly misclassified, outcome variable is a particular function of the model's explanatory variables. Our proposed method does not require the strong identifying assumptions of measurement error problems essentially because error-free measurement is available from a clerical sample which identifies correct links. The assumptions we make in this paper are outlined in section 3.

Section 2 summarises the ML approach to contingency table and regression analysis under perfect linkage. Section 3 considers the ML approach in the presence of incorrect links. Section 4 considers the ML approach in the presence of both incorrect links and unlinked records. Section 5 demonstrates the effectiveness of many of the proposed estimators in an empirical study. Section 6 summarises the findings.

2. Perfect linkage

By way of introducing notation, this section discusses the case where the linkage is perfect. The estimating approach in this section is standard since, clearly, no special adjustment for incorrect linkage is required. Section 2.1 discusses estimating cell probabilities in a contingency table and section 2.2 discusses estimating regression coefficients in a logistic regression.

2.1 Contingency tables

For notation, it is convenient when considering contingency table analysis to transform \mathbf{x}_i to a single categorical variable x so that $x = 1, 2, \dots, g, \dots, G$. Define y to be a categorical variable on file Y, where $y = 1, \dots, c, \dots, C$.

Consider the following factorisation of the distribution of x and y

$$p(y, x) = p_1(y | x; \Pi) p_2(x),$$

where $\Pi = (\pi'_1, \dots, \pi'_g, \dots, \pi'_G)'$, $\pi_g = (\pi_{1|g}, \dots, \pi_{c|g}, \dots, \pi_{C|g})'$, $\pi_{c|g}$ is the probability that $y = c$ given $x = g$. We assume that for every value of x there are C possible values of y which implies that the dimension of Π is CG .

We now consider maximum likelihood estimation of the parameter Π , characterising p_1 , under perfect linkage. Perfect linkage means that all records on file X are correctly linked to their corresponding record on file Y (*i.e.*, there are no incorrect links and no unlinked records). Under perfect linkage, $n_{xy} = n_x$ and the set of linked records is denoted by $\mathbf{d} = \{(y_i, x_i): i = 1, \dots, n_{xy}\}$. Under perfect linkage, the score function for $\pi_x = (\pi_{1|x}, \dots, \pi_{c|x}, \dots, \pi_{C|x})'$ characterised by the multinomial distribution, is

$$\text{Score}(\pi_x; \mathbf{d}) =$$

$$(\text{Score}(\pi_{1|x}; \mathbf{d}), \dots, \text{Score}(\pi_{c|x}; \mathbf{d}), \dots, \text{Score}(\pi_{C-1|x}; \mathbf{d}))' \quad (1)$$

where

$$\begin{aligned} \text{Score}(\pi_{c|x}; \mathbf{d}) &= \sum_i (w_{ic|x} \pi_{ic|x}^{-1} - w_{ic|x} \pi_{ic|x}) \\ &= n_{c|x} \pi_{c|x}^{-1} - n_{c|x} \pi_{c|x}, \end{aligned}$$

for $c = 1, \dots, C-1$, where $n_{c|x} = \sum_i w_{ic|x}$, $w_{ic|x} = 1$ if $y_i = c$ and $x_i = x$ and $w_{ic|x} = 0$ otherwise, and the category corresponding to $y = C$ is the arbitrarily chosen reference category. Solving $\text{Score}(\pi_x; \mathbf{d}) = \mathbf{0}_{C-1}$ for π_x , where $\mathbf{0}_{C-1}$ is a $C-1$ column vector of zeros, gives the maximum likelihood (ML) estimator

$$\hat{\pi}_{c|x} = n_{c|x} / n_x, \quad (2)$$

where

$$n_x = \sum_i \sum_c w_{ic|x}$$

and

$$\hat{\pi}_{C|x} = 1 - \sum_{c=1}^{C-1} \hat{\pi}_{c|x}.$$

2.2 Logistic regression

Consider the logistic regression model

$$E(y_i) = v_i \quad (3)$$

$$v_i = 1 / [1 + \exp(\beta' \mathbf{x}_i)]. \quad (4)$$

For (4) the K elements of \mathbf{x}_i are dichotomous variables and y_i is now a dichotomous variable available from File Y. If we define $\mathbf{x} = (\mathbf{x}_1, \dots, \mathbf{x}_i, \dots, \mathbf{x}_{n_y})'$, $\mathbf{y} = (y_1, \dots, y_i, \dots, y_{n_y})'$ and $\mathbf{v} = (v_1, \dots, v_i, \dots, v_{n_y})'$, the score matrix for β based on perfectly linked data, \mathbf{d} , is

$$\text{Score}(\beta; \mathbf{d}) = \mathbf{x}'(\mathbf{y} - \mathbf{v}). \quad (5)$$

Solving $\text{Score}(\beta; \mathbf{d}) = \mathbf{0}_K$ for β gives the ML estimate $\hat{\beta}$, which can be found by applying the well-known Newton-Raphson method.

3. Analysis with incorrect links

This section considers the situation where the linked file contains incorrect links but does not contain unlinked records. This occurs when all the records on File X are linked to a record on File Y (so $n_x \leq n_y$). Define the linked file of records by $\mathbf{d}^* = \{d_i^* = (y_i^*, x_i): i = 1, \dots, n_x\}$, where y_i^* is the value of y that is linked to record i on file X. To clarify, y_i is the true value of y for record i on file X, so that $y_i^* = y_i$ if record i is correctly linked.

The estimator given by (2), together with the assumption that $y_i^* = y_i$ for $i = 1, \dots, n_x$, is naive since it treats the probabilistically linked file as if it were perfectly linked. In general the naive estimator will be biased. This section derives ML estimators which account for the fact that the data have been linked probabilistically or linked imperfectly in some way.

It is common practice to select a subsample of the linked file, denoted by s_c , which is then reviewed clerically. The clerical review classifies a link, \mathbf{d}_i , as either correct or incorrect. Let $\delta_i = 1$ if record i on File X is correctly linked and $\delta_i = 0$ otherwise.

Designing the clerical subsample is an important problem, especially since clerical review is often a costly exercise. Possible uses of a clerical sample include estimating the proportion of correctly linked and unlinked records, to assist in deciding which records should be linked and which should remain unlinked, to ensure correct inference using \mathbf{d}^* (i.e., the purpose of this paper), and to identify improvements to the way in which records are linked (in the ABS applications mentioned above, clerical samples were designed to ensure that each link had at least a specific probability of being correct). For the purpose of making correct inference using \mathbf{d}^* selecting the clerical sample by simple random sampling is a reasonable approach. A more efficient clerical subsample could possibly be devised but there is no obvious way to do so. This is because the parameters that we need to estimate to implement the ML method described in this paper depend upon the specific analysis (e.g., choice of y and \mathbf{x}). Designing a clerical sample for all possible analyses would be difficult.

We factorise the joint distribution $p(y_i, \mathbf{x}_i, \delta_i)$ by

$$p(y_i | \mathbf{x}_i; \boldsymbol{\theta}) p(\mathbf{x}_i) p(\delta_i | \mathbf{x}_i), \quad (6)$$

where $\boldsymbol{\theta} = \boldsymbol{\beta}$ in the regression case, $\boldsymbol{\theta} = \boldsymbol{\Pi}$ in the contingency table case. Factorisation (6) means that the links are incorrect at random (IAR) or, in other words, that the distributions $y_i | \mathbf{x}_i$ and $\delta_i | \mathbf{x}_i$ are independent. Under this assumption it is only necessary to maximise the likelihood associated with the factor $p(y_i | \mathbf{x}_i; \boldsymbol{\theta})$. Throughout this section we assume (6). It is important to point out that (6), and the development that follows, makes no assumption requiring File X to be a subset of File Y (e.g., when units on File X are a subsample of the units on File Y) or that the linkage process involves a single pass. We also assume that the correctness of linkage, δ_i , is independent from record to record.

As mentioned in the introduction, each linked record is assigned a score based on the probability that the records belong to the same unit. Denote the score by r_i . A referee suggested using r_i to more accurately parameterise the distribution of δ_i . Technically this suggestion would

involve replacing $p(\delta_i | \mathbf{x}_i)$ with $p(\delta_i | \mathbf{x}_i, r_i)$ in (6) and would likely reduce the variability of the ML estimators discussed in section 3. This would be a useful avenue of further research.

3.1 Contingency tables

Define $w_{ic|x}^* = 1$ if $y_i^* = c$ and $x_i = x$, and $w_{ic|x}^* = 0$ otherwise. The expectation of $w_{ic|x}^*$ given \mathbf{d}_i^* is

$$\begin{aligned} E_{\mathbf{d}^*}(w_{ic|x}^* | x_i = x, y_i^* = y^*) &= \\ &= w_{ic|x}^* p_{xy^*} + (1 - p_{xy^*}) \pi_{c|x} \quad \text{if } i \notin s_c \\ &= w_{ic|x}^* \quad \text{if } i \in s_c \text{ and } \delta_i = 1 \\ &= \pi_{c|x} \quad \text{if } i \in s_c \text{ and } \delta_i = 0 \end{aligned}$$

and p_{xy^*} is the probability that the i^{th} link is correct given $x_i = x$ and $y_i^* = y^*$. The ML estimator of $\pi_{c|x}$ using the probabilistically linked data, \mathbf{d}_i^* , is then

$$\tilde{\pi}_{c|x} = \tilde{n}_{c|x} \left(\sum_c \tilde{n}_{c|x} \right)^{-1} \quad (7)$$

where

$$\tilde{n}_{c|x} = \sum_i \tilde{w}_{ic|x}, \quad (8)$$

$$\begin{aligned} \tilde{w}_{ic|x} &= w_{ic|x}^* \hat{p}_{xy^*} + (1 - \hat{p}_{xy^*}) \tilde{\pi}_{c|x} \quad \text{if } i \notin s_c \\ &= w_{ic|x}^* \quad \text{if } i \in s_c \\ &= \tilde{\pi}_{c|x} \quad \text{if } i \in s_c \text{ and } \delta_i = 0 \end{aligned} \quad (9)$$

and

$$\hat{p}_{xy^*} = \left(\sum_{i \in s_c} w_{ic|x}^* \delta_i \right) \left(\sum_{i \in s_c} w_{ic|x}^* \right)^{-1}. \quad (10)$$

The estimation procedure involves iterating between (7), (8) and (9) until convergence. Specifically the algorithm is:

1. Calculate \hat{p}_{xy^*} from (10).
2. Initialise $\tilde{\pi}_{c|x}^{(0)}$ and then calculate $\tilde{w}_{c|x}^{(0)}$ from (9) and then $\tilde{n}_{c|x}^{(0)}$ from (8).
3. Calculate $\tilde{\pi}_{c|x}^{(t)}$ from (7) using $\tilde{n}_{c|x}^{(t-1)}$.
4. Calculate $\tilde{w}_{c|x}^{(t)}$ from (9) using $\tilde{\pi}_{c|x}^{(t)}$ and then calculate $\tilde{n}_{c|x}^{(t)}$ from (8) using $\tilde{w}_{c|x}^{(t)}$.
5. Iterate between 3 and 4 until convergence.

The initialised value $\tilde{\pi}_{c|x}^{(0)}$ could be set to the naive estimate of $\pi_{c|x}$, which was described in section 3 above. However, our experience was that the choice of initial value was not important.

3.2 Logistic regression

Below we describe two ML methods (Methods 1 and 2) for estimating $\boldsymbol{\beta}$ using the probabilistically linked data, \mathbf{d}^* .

Both methods give unbiased estimates under the IAR assumption. The difference between the methods is the level of aggregation at which the probabilities of correct linkage are estimated. Method 1 requires these probabilities at a fine level of aggregation, which may mean its estimates are more variable than those of Method 2.

3.2.1 Method 1

The expectation of y conditional on the linked data is

$$\begin{aligned} E_{d^*}(y_i | \mathbf{x}_i = \mathbf{x}, y_i^* = y^*) &= \\ y_i^* p_{xy^*} + (1 - p_{xy^*}) v_i &\text{ if } i \notin s_c \\ &= y_i^* \text{ if } i \in s_c \text{ and } \delta_i = 1 \\ &= v_i \text{ if } i \in s_c \text{ and } \delta_i = 0 \end{aligned}$$

and p_{xy^*} is the probability that the i^{th} link is correct given $x = x_i$ and $y_i^* = y^*$.

The ML estimator is then obtained by iterating between finding the solution, denoted by $\tilde{\beta}$, for β in (5) with y_i replaced by \tilde{y}_i , where

$$\begin{aligned} \tilde{y}_i &= y_i^* \hat{p}_{xy^*} + (1 - \hat{p}_{xy^*}) \tilde{v}_i \text{ if } i \notin s_c \\ &= y_i^* \text{ if } i \in s_c \text{ and } \delta_i = 1 \\ &= \tilde{v}_i \text{ if } i \in s_c \text{ and } \delta_i = 0, \end{aligned} \quad (11)$$

\tilde{v}_i has the same form as v_i except that β is replaced with $\tilde{\beta}$ and \hat{p}_{xy^*} is the estimated proportion of correct links in the clerical sample for each combination of \mathbf{x} and y^* .

3.2.2 Method 2

Let $\mathbf{x}'\mathbf{y}$ in (5) have k^{th} element

$$r_k = \mathbf{x}'_k \mathbf{y} = \sum_i y_i x_{ik} = \sum_i r_{ik},$$

where $r_{ik} = y_i x_{ik}$. The expectation of r_{ik} conditional on \mathbf{d}^* is

$$\begin{aligned} E_{d^*}(r_{ik} | \mathbf{x}_i = \mathbf{x}, y_i^* = y_i^*) &= \\ [y_i^* p_{ky^*} + (1 - p_{ky^*}) v_i] x_{ik} &\text{ if } i \notin s_c \\ &= y_i^* x_{ik} \text{ if } i \in s_c \text{ and } \delta_i = 1 \\ &= v_i x_{ik} \text{ if } i \in s_c \text{ and } \delta_i = 0 \end{aligned} \quad (12)$$

and p_{ky^*} is the probability that a link with $x_{ik} = 1$ is correct given $y_i^* = y^*$. The ML estimator is then obtained by iterating between finding the solution, denoted by $\tilde{\beta}$, for β in (5) with r_{ik} replaced by \tilde{r}_{ik} , where

$$\begin{aligned} \tilde{r}_{ik} &= [y_i^* \hat{p}_{ky^*} + (1 - \hat{p}_{ky^*}) \tilde{v}_i] x_{ik} \text{ if } i \notin s_c \\ &= y_i^* x_{ik} \text{ if } i \in s_c \text{ and } \delta_i = 1 \\ &= \tilde{v}_i x_{ik} \text{ if } i \in s_c \text{ and } \delta_i = 0, \end{aligned} \quad (13)$$

\tilde{v}_i has the same form as v_i except that β is replaced with $\tilde{\beta}$ and \hat{p}_{ky^*} is the estimated proportion of correct links in the clerical sample for each combination of \mathbf{x} and y^* . Namely, if $y_i^* = 1$,

$$p_{ky^*} = \left(\sum_{i \in s_c} y_i^* x_{ik} \delta_i \right) \left(\sum_{i \in s_i} y_i^* x_{ik} \right)^{-1}$$

and if $y_i^* = 0$,

$$p_{ky^*} = \left(\sum_{i \in s_c} (1 - y_i^*) x_{ik} \delta_i \right) \left(\sum_{i \in s_c} (1 - y_i^*) x_{ik} \right)^{-1}.$$

This approach requires only $2K$ probabilities to be calculated from the clerical sample and, on this basis, may be preferable to the approach in section 3.2.1 which requires more probabilities to be calculated.

3.3 Estimating the variance using the bootstrap

In this section we describe how to calculate the variance of the ML estimates of section 3. Denote the parameter of interest by θ , introduced earlier, and its ML estimate by $\tilde{\theta}$. The Bootstrap (Rubin and Little 2003) estimate of the variance of $\tilde{\theta}$, denoted by $\hat{v}_{\text{boot}}(\tilde{\theta})$, is obtained by

1. Taking a replicate sample of size n_x from the linked file, \mathbf{d}^* , by simple random sampling with replacement. Denote the r^{th} replicate sample by $\mathbf{d}^*(r)$. The r^{th} replicate clerical sample is $s_c(r) = s_c \cap \mathbf{d}^*(r)$.
2. Calculating $\tilde{\theta}(r)$ which has the same form as $\tilde{\theta}$ except that $\mathbf{d}^*(r)$ is used instead of \mathbf{d}^* and $s_c(r)$ is used instead of s_c .
3. Repeating steps 1 and 2 R times, where R is the number of replicates.
4. Calculating

$$\hat{v}_{\text{boot}}(\tilde{\theta}) = \frac{1}{R} \sum_{b=1}^R (\tilde{\theta}(b) - \tilde{\theta})(\tilde{\theta}(b) - \tilde{\theta})'.$$

4. Analysis with incorrect links and unlinked records

This section discusses two ways of analysing linked data in the presence of incorrect links and unlinked records. As mentioned in the introduction, the problem of analysis when there are unlinked records has clear parallels with the problem of unit non-response. Unlinked records may result in some characteristics on the linked file being over- or under-represented, thus leading to biased analysis. As discussed in more detail below, we use the fact that the mechanism giving rise to unlinked records can only be a function of \mathbf{z} .

This section considers two methods of making inference in the presence of incorrect links and unlinked records, where linked records are indexed by $i = 1, \dots, n^*$. (Remember that the i^{th} record on File X is an *unlinked record* if $i \in U_{xy}$ and record i was not linked to any record on File Y.) The methods involve independently modelling the processes that determine which records are incorrectly linked and which are unlinked (see section 5 for an illustration). These models require a subsample, denoted by s_{xc} , of all records on File X to be subjected to clerical review. Records in the subsample will be either linked to records on File Y or not linked. Linked records in the subsample must be identified as either correctly or incorrectly linked by the clerical review process. A subsample record which is not linked must be identified as either *unlinked*, or *otherwise*. *Unlinked* means the corresponding record was found on File Y but not linked to it, whereas *otherwise* indicates the corresponding record was not found on File Y and therefore assumed not to exist. The latter identification is potentially much more difficult and time-consuming than the former because it assumes some other error-free process is available for checking whether links, which were not made, are in fact correct. Unlinked records, by their nature, have limited information that can be used to identify the correct link, even during clerical review. Such a process may not exist, in which case adjusting for unlinked records would seem to be impossible. However, such a process may involve a clerical review of names appearing on the two files to be linked. For example, a clerical reviewer may realise that the names *John O. Smith* and *Joh O. Smith* on two different records may in fact be the same name (with an “n” missing in the latter case, perhaps due to errors in scanning), whereas the automated linking process may treat the two names as completely different. The clerical reviewer may then decide that the above two records correspond to the same individual and so therefore should be linked. (Bishop (2009) and Wright (2009) discuss the benefits of clerical review).

The first method involves conditioning analysis on a variable $\zeta_i = \zeta_i(\mathbf{z}_i)$. The variable ζ is defined so that inference, in the presence of unlinked records, is unbiased conditional on ζ . The term ζ is introduced since, in many cases, it would be impractical or unnecessary to condition on all the information in \mathbf{z} . It is possible to give ζ_i a non-missing value even when \mathbf{z}_i contains missing values. The exact form of the function $\zeta(\mathbf{z})$ would need to be justified after analysis of the subsample, s_{xc} . For example, if persons under 20 years of age are under-represented in the linked file, ζ would indicate whether a person is under 20 years of age. One approach to analysis is to include ζ as a covariate in the regression model. The method in section 3 would then apply directly. However, analysts may like to integrate over ζ so that it does not appear in the logistic model or

contingency table. Section 4.2 discusses how to do this for contingency tables. Section 4.3 discusses a pseudo-likelihood approach which assigns weights to the linked records that attempt to account for any under- or over-representation of certain subpopulations in the linked data. Again, the choice of weight would need to be justified after analysis of the subsample, s_{xc} , which identifies unlinked records. This is discussed further in the context of the empirical study.

4.1 Can we ignore unlinked records?

Define the variable $\gamma_i = 1$ if record i on File X is unlinked and $\gamma_i = 0$ otherwise. Also let ζ_i be a variable so that $\zeta_i = 1, 2, \dots, h, \dots, H$, where H is the number of categories for ζ . We can ignore the fact that there are unlinked records if we are prepared to assume that, conditional on \mathbf{x}_i , the distributions of y_i , γ_i and δ_i are independent. Technically this assumption leads to the factorisation,

$$p(y_i, \mathbf{x}_i, \delta_i, \gamma_i, \zeta_i) \propto p(y_i | \mathbf{x}_i; \boldsymbol{\theta}) p(\delta_i | \mathbf{x}_i) p(\gamma_i | \mathbf{x}_i) p(\zeta_i)$$

where again $\boldsymbol{\theta} = \boldsymbol{\beta}$ or $\boldsymbol{\Pi}$. It is worthwhile checking whether this assumption is valid from the clerical subsample. If the assumption is reasonable, then there is no need to apply the methods in section 4.2 and 4.3 and the methods in section 3 will suffice.

We may not be prepared to make the assumption mentioned above. We may however be prepared to assume, conditional on \mathbf{x} and ζ , the distributions of y_i , γ_i and δ_i are independent. In this case, we say unlinked records are not ignorable. Technically this assumption leads to the factorisation,

$$p(y_i, \mathbf{x}_i, \delta_i, \gamma_i, \zeta_i) \propto p(y_i | \mathbf{x}_i, \zeta_i; \boldsymbol{\Lambda}) p(\delta_i | \mathbf{x}_i; \boldsymbol{\tau}) p(\gamma_i | \mathbf{x}_i, \zeta_i) p(\zeta_i)$$

where $\boldsymbol{\Lambda}$ is the parameter for the distribution of $y_i | \mathbf{x}_i, \zeta_i$. If we are interested in $p(y_i | \mathbf{x}_i; \boldsymbol{\theta})$ but not $p(y_i | \mathbf{x}_i, \zeta_i; \boldsymbol{\Lambda})$, one approach is to integrate out (*i.e.*, average over) ζ_i from the latter.

4.2 Conditional Maximum Likelihood (CML) for contingency tables

First, parameterise the joint distribution of y_i , x_i and ζ_i by the multinomial distribution with parameter, $\boldsymbol{\Lambda}$. Define $\boldsymbol{\Lambda} = (\boldsymbol{\Pi}'_1, \dots, \boldsymbol{\Pi}'_h, \dots, \boldsymbol{\Pi}'_H)'$, where $\boldsymbol{\Pi}_h = (\boldsymbol{\pi}'_{1h}, \dots, \boldsymbol{\pi}'_{gh}, \dots, \boldsymbol{\pi}'_{Gh})'$, $\boldsymbol{\pi}_{gh} = (\pi_{1|gh}, \dots, \pi_{c|gh}, \dots, \pi_{C|gh})'$ and $\pi_{c|gh}$ is the probability that $y_i = c$, $x_i = g$ and $\zeta_i = h$. The ML estimator of $\boldsymbol{\Pi} = (\pi_{c|x})$ from section 2.1 when linkage errors are not ignorable is $\tilde{\boldsymbol{\Pi}} = (\tilde{\pi}_{c|x})$, where

$$\tilde{\pi}_{c|x} = \sum_{h=1}^H \tilde{\pi}_{c|xh} \hat{\pi}_{h|x}, \quad (14)$$

where

$$\tilde{\pi}_{c|xy} = \tilde{n}_{c|xy} \left(\sum_c \tilde{n}_{c|xy} \right)^{-1}, \quad (15)$$

$\tilde{n}_{c|xy} = \sum_{i \in U_i} \tilde{w}_{ic|xy}$, $\sum_{i \in U_i}$ is the sum over the n^* linked records and $\tilde{\pi}_{h|x}$ for $h = 1, \dots, H$ is the standard estimate of the marginal distribution of ζ given x on File X. Further, if $i \notin S_c$

$$\tilde{w}_{ic|xy} = w_{ic|xy}^* \hat{p}_{xy^*h} + (1 - \hat{p}_{xy^*h}) \tilde{\pi}_{c|xy}, \quad (16)$$

\hat{p}_{xy^*h} is the probability that the i^{th} link is correct given $x_i = x$, $\zeta_i = h$ and $y_i^* = y^*$, $w_{ic|xy}^* = 1$ if $x_i = x$, $\zeta_i = h$ and $y_i^* = y^*$, and $w_{ic|xy}^* = 0$ otherwise. If $i \in S_c$, then $\tilde{w}_{ic|xy} = w_{ic|xy}^* = w_{ic|xy}$ if the link is determined to be correct and $\tilde{w}_{ic|xy} = \tilde{\pi}_{c|xy}$ if it is determined to be incorrect.

The ML estimator $\tilde{\pi}_{c|x}$ is obtained by iterating between (14), (15) and (16) until convergence.

4.3 Pseudo-Maximum Likelihood (PML)

This section discusses an alternative to the CML, discussed in section 4.2, which is referred to as Pseudo-Maximum Likelihood (see Chambers and Skinner 2003). It is essentially a weighting approach, which may be easier to implement than CML, and relies on the factorisation given in section 4.2. It involves solving weighted versions of the score functions, $\text{Score}(\pi_x; \mathbf{d}) = \mathbf{0}_{C-1}$ and $\text{Score}(\beta; \mathbf{d}) = \mathbf{0}_K$ for π_x and β respectively, where a record's weight equals the inverse of the probability that the record will remain unlinked. We denote the probability that record i will not remain unlinked by $t_i = E(\gamma_i)$ so that the unit weights are given by $q_i = t_i^{-1}$, where here $i = 1, \dots, n^*$. Consequently the PML estimator for $\pi_{c|x}$ is

$$\tilde{\pi}_{c|x}^{\text{PML}} = \tilde{n}_{c|x} \left(\sum_c \tilde{n}_{c|x} \right)^{-1}, \quad (17)$$

where $\tilde{n}_{c|xy} = \sum_{i \in U_i} q_i \tilde{w}_{ic|xy}$. The estimate of $\tilde{\pi}_{c|x}^{\text{PML}}$ is obtained by iterating between updating $\tilde{w}_{ic|x}$, given by (7), and (17) until convergence. The PML estimator for β is the same as the ML estimator but where the estimating equation (5) now has unit weights of q_i . One possible approach to estimating the accuracy of the PML estimates under perfect linkage is to use the Bootstrap method as described earlier, but where now the weight q_i is introduced.

To illustrate when unlinked records are not ignorable, consider linking a data base with personal employment status to another data base with education level. Also assume that age and sex variables, which are correlated with employment and education, are available on one of the data bases. After conducting a clerical review, we may find that

records for young males are 50% more likely to remain unlinked than records for females. This could be because males are less likely to provide their personal information, which is useful in linkage. Clearly, records for males on the linked file need to be given a weight double that for females in order for joint analysis of employment status and educational level to be unbiased.

5. Empirical study

A quality study conducted by the Australian Bureau of Statistics involved linking the 2006 Census of Population and Housing to its Dress Rehearsal. The Census Dress Rehearsal collected information from 78,349 persons and was conducted one year before the Census. The 2006 Census collected information from more than 19 million people.

Within a short window, during which the 2006 Census data were being processed, name and address were available for both the Census and the Census Dress Rehearsal. During this time, the two files of person level records were linked using two different standards of information:

- *Gold Standard* (GS) used name, address, mesh block and selected Census data items. Mesh block is a geographic area typically containing 50 dwellings. All names and addresses were destroyed at the end of the Census processing period.
- *Bronze Standard* (BS) used mesh block and selected Census data items (*i.e.*, did not use name and address). This is a method proposed to be used for future linking work by the ABS.

Full details of the quality study and the linkage methodology are given in Solon and Bishop (2009). The role of GS in the quality study is critical. It provides a benchmark against which the reliability of BS can be compared. The usefulness of the GS as a benchmark is due to the fact that name and address are powerful variables for the purpose of identifying common individuals on the Census and CDR and that it was subjected to thorough clerical review. As a result, GS is assumed to correspond to perfect linkage. Accordingly, differences between estimates based on GS and BS are interpreted as error. In other words, interest focuses on the reliability of BS *relative* to GS.

5.1 Linking methodology

5.1.1 Blocking and linking variables and the 1 – 1 assignment algorithm

This subsection provides an overview of the CDR-to-Census linkage methodology for BS. The linking method consisted of a sequence of passes, where each pass is

defined by a set of blocking and linking variables and a 1 - 1 assignment algorithm. In the case of multiple passes, only records not linked in the first pass are eligible to be linked in the second pass, and only records not linked in the second pass are eligible to be linked in the third pass, and so on.

Table 1 gives the blocking variables, denoted by “B” for the BS. For example, during Pass 1, a Census record and a CDR record are only considered as a possible link if they have the same value for mesh block.

Linking variables are used to measure the degree of agreement between a record pair. A high level of agreement suggests that the likelihood of the record pair constituting a correct link is high. Table 1 gives the linking variables, denoted by “L”, for BS. For example, during Pass 1 of BS, a range of variables such as day, month and year of birth, country of birth and highest level of qualifications are used as linking variables.

Table 1

An example of blocking (B) and linking (L) variables used when linking 2006 Census data with the Census Dress Rehearsal. Different blocking variables were used on each of the two passes

Variable	Pass 1	Pass 2
Day of birth	L	B
Month of birth	L	B
Year of birth	L	B
Sex	L	B
Indigenous status	L	L
Country of birth	L	L
Language spoken	L	L
Year of arrival	L	L
Marital status	L	L
Religious affiliation	L	L
Field of study of highest qualification	L	L
Level of highest qualification	L	L
Highest level of schooling	L	L
Mesh block	B	L

An output from each pass is a score for all record pairs. The score is a measure of the level of agreement between the pair of records. We defer the formal definition of score (for details see (3.6), Conn and Bishop 2006) but illustrate how it can be interpreted below. Consider BS in Pass 2 where record pairs have the same full date of birth and sex; a record pair would be assigned a score of 23.5 if there is agreement on mesh block (+17) and year of arrival (+8) and disagreement on religion (-1.5) (in this example agreement status for other linking variables would contribute to the score but for illustration purposes we ignore them). The contribution to the score for agreement on mesh block (+17) is greater than that for agreement on year of arrival (+8) because the former is less likely to occur by chance alone.

To formalise the aim of the linkage algorithm, denote the score for record i on the CDR and record j on the Census

during pass p of BS by r_{pij} . The set of all record pair scores r_{pij} and the cut-off f_p were used by the linking package *Febri* (see Christen and Churches 2005) to determine the optimal set of links in pass p . The term f_p is the minimum value for the score in order for a record pair to be assigned as a link during pass p . The *Febri* algorithm seeks to maximise $\sum_i r_{pij}$, subject to $r_{pij} > f_p$. Clearly, the number of links depends upon f_p .

In what follows, we evaluate BS with two different sets of cut-offs, where a set of cut-offs is defined by the pass 1 and 2 cut-offs. The first is referred to as the Very Low (VL) cut-off and is considered to be optimal cut-off since, for a range of cut-offs, its naive estimates were “closest” to the corresponding GS estimates (see Bishop 2009). The second cut-off is referred to as Ultra-Low (UL) and effectively seeks to maximise the number of linked CDR records. Below we refer to the two BS linked files by their cut-offs, VL and UL.

5.1.2 Linking results

GS linked 70,274 of the 78,349 CDR records. Under the assumption that GS corresponds to perfect linkage, there were 8,075 individuals with CDR records but no Census records. In reality the GS is not perfect. For a discussion on this see Bishop 2009.

VL linked 57,790 CDR records. Of the 70,274 CDR records that were linked by GS, 13,784 remained unlinked by VL, 700 were linked incorrectly by VL and 55,790 were linked correctly by VL. Also, 1,300 CDR records were linked by VL but were not linked by GS- these are also incorrect links. So in total there were 2,000 (= 700 + 1,300) incorrect links.

UL linked 74,350 CDR records. Of the 70,274 CDR records that were linked by GS, 2,811 remained unlinked by UL, 9,793 were linked incorrectly by UL and 57,670 were linked correctly by UL. Also, 6,887 CDR records were linked by UL but were not linked by GS.

In summary, 97% of the VL links are correct and 20% (= 13,784/70,274) of the GS' CDR records remain unlinked. The corresponding figures for UL are 78% and 4% (= 2,811/70,274).

5.1.3 Modelling the probability of a link being correct

All UL and VL links were known to be correct or incorrect (e.g., if a UL link is also made by GS then the UL link is correct. Otherwise the UL link is incorrect). As a result, p_{xy} , in section 3.1 was known from GS. However, to simulate reality, p_{xy} was estimated from a clerical sample of size 1,000 that was selected from the linked files by simple random sampling.

5.1.4 Modelling the probability of a record remaining unlinked

Each CDR record linked by the GS was assigned a variable which indicated whether the record was unlinked by BS. Namely, if the record remained unlinked by BS then the indicator variable was assigned a '1' otherwise a '0'. A logistic model was fitted using GS, where the response variable was the above indicator variable and the explanatory variables were obtained from the CDR. The more than 20 explanatory variables that are in the model were selected by standard forward-backward model selection. The explanatory variables included educational level, language, born overseas, Indigenous status, and indicators of missing key variables such as *meshblock*. The resulting prediction resulted in t_i and was used below to implement the Pseudo-ML method for both contingency tables and logistic regression.

5.2 Results of tabular analysis

Table 2 gives the results of cross-tabulating employment status of indigenous people as reported on the CDR and Census. Table 2a shows that the GS estimate of the proportion of indigenous people employed in the Census, given they were employed in CDR, is 78.3%. The corresponding naive estimate for VL, which assumes the data are perfectly linked, is 86.7%. Even after replacing each of the 700 incorrect VL links by their corresponding correct link and discarding the 1,300 linked records for which no correct link exists, the naive estimate is largely unchanged at 86.0% (referred to as *Gold Links* in Table 2a). This shows that the difference between the VL and GS estimates is not so much due to incorrect links but is mainly due to unlinked records. This explains in part why the ML estimate (86.4%) for VL (see section 3.1), which only corrects for incorrect links, did not lead to much improvement. Conditional ML (CML) (see

section 4) was considered in an attempt to reduce the error due to unlinked records that may have led to a misrepresentation, with respect to age and sex characteristics, in the linked file. The CML employment estimate was 86.6%. Unfortunately, CML did not make much of an improvement, indicating that the underlying mechanism generating unlinked records did not depend upon age and sex. PML estimates (see section 4) also did not make much of an improvement, indicating that the logistic model described in section 5.1.4 did not explain the mechanism generating unlinked records. Interestingly, the ML estimate using UL was 81.8%—by far the closest estimate to the GS estimate of 78.3%. The UL's main source of error is due to incorrect links, the type of linkage error which the ML estimator addresses. This indicates that correcting for errors due to incorrect links was much more successful than correcting for errors due to unlinked records.

Standard errors of the GS, naive and ML estimates are shown in parentheses in Table 2a. For VL and UL, ML standard errors are respectively about 25% and 75% larger than the corresponding naive standard errors. Also, the ML standard errors for UL are slightly smaller than for VL indicating that the extra links made by UL were worthwhile. Clearly, naive inference with UL over-states the level of confidence in estimates. For VL, naive and ML standard errors and estimates are very close.

Irrespective of the cut-off, the ML estimates in Table 2 a, b and c are always closer to the GS estimates than the corresponding naive estimate. For example in Table 2b the ML estimates for VL is 36.9%, noticeably closer to the GS estimate of 37.9% than the naive estimate of 33.3%. Based on the estimates in Table 2 it could be argued that the choice of whether to use VL or UL is not so important, as long as the ML estimator is used.

Table 2
Percentages of Indigenous persons in various employment categories in 2006 given their employment category in 2005. For each linked data set, Very Low and Ultra Low, the estimation methods can be compared with the Gold

Estimates for different methods and linked data set								
a: Indigenous persons employed in 2005								
Status in 2006	Gold		Gold links	Very Low Cut-off			Ultra Low Cut-off	
		Naive		ML	PML	CML	Naive	ML
Employed	78.3 (1.7)	86.7 (2.4)	86.0	86.4 (3.0)	86.6	86.1	71.9 (1.7)	81.8 (2.9)
Unemployed	3.7 (0.84)	4.2 (1.2)	4.3	4.1 (2.5)	4.1	4.2	6.3 (0.82)	3.3 (2.1)
Not in the labour force	17.8 (1.6)	9.0 (2.4)	9.6	9.3 (3.1)	9.1	9.6	21.6 (1.6)	14.7 (2.8)
b: Indigenous persons unemployed in 2005								
Status in 2006	Gold		Naive	Very Low		Naive	Ultra Low	
				ML			ML	
Employed	27.5		27.7	27.2		35.2	23.8	
Unemployed	34.4		38.9	36.4		32.3	38.0	
Not in the labour force	37.9		33.3	36.3		32.3	38.0	
c: Indigenous persons not-in-the-labour force in 2005								
Employed	13.7		10.8	10.7		24.3	10.5	
Unemployed	5.8		7.6	7.4		6.3	5.8	
Not in the labour force	80.4		81.5	81.8		69.2	83.5	

Table 3 is the same as Table 2 except that it describes analyses of linked records from all persons 15 and over rather than only Indigenous persons. Again the ML always makes an improvement for the UL, though this is not the case for VL. Table 4 gives the student status in 2006 for persons who were students in 2005. Again the ML generally makes the estimates closer to the corresponding Gold estimate, especially for UL.

Table 3

Percentages of all persons aged over 15 in various employment categories in 2006 given their employment category in 2005. For each linked data set, Very Low and Ultra Low, the estimation methods can be compared with the Gold

Estimates for different methods and linked data set					
Status in 2006	Gold	Very Low		Ultra Low	
		Naive	ML	Naive	ML
a: Persons employed in 2005					
<i>Employed</i>	91.8	92.2	92.6	89.7	92.4
<i>Unemployed</i>	1.8	1.7	1.6	1.9	1.6
<i>Not in the labour force</i>	6.2	6.1	5.6	8.3	5.8
b: Persons unemployed in 2005					
<i>Employed</i>	44.5	44.3	44.0	49.4	43.8
<i>Unemployed</i>	26.8	26.6	27.5	22.8	27.6
<i>Not in the labour force</i>	28.6	28.7	28.4	27.6	28.5
c: Persons not-in-the-labour force in 2005					
<i>Employed</i>	12.1	12.3	11.1	16.8	11.0
<i>Unemployed</i>	3.1	3.1	3.0	3.0	3.0
<i>Not in the labour force</i>	84.7	84.5	85.7	80.1	85.9

Table 4

Student outcomes in 2006 for high school students in 2005

Student Status in 2006	Gold	Very Low		Ultra Low	
		Naive	ML	Naive	ML
<i>High School Student</i>	79.3	79.3	79.6	77.4	79.6
<i>Completed High School</i>	14.0	14.3	13.7	14.7	14.1
<i>Did not Complete High School</i>	6.6	6.3	6.6	7.8	6.2

5.3 Simulation

The following simulation study illustrates the problems with naive analysis and the benefit of using the method outlined in this paper. Files X and Y in the simulation, each containing 2,000 records, are independently generated 400 times, where each generated file is denoted by $X(r)$ and $Y(r)$, and $r = 1, \dots, 400$. Specifically, on $X(r)$ x_i is randomly generated from the Bernoulli distribution with parameter 0.5. On $Y(r)$, y_i is randomly generated from the

Bernoulli distribution with parameter v_i , where $v_i = 1 / [1 + \exp(\beta_0 + \beta_1 x_i)]$, $\beta = (\beta_0, \beta_1)'$, $\beta_0 = -0.5$, $\beta_1 = 1.5$. The r^{th} set of imperfectly linked data, $\mathbf{d}^*(r)$, is generated by correctly linking each record on File $Y(r)$ to one record on File $X(r)$ with probability $p = 0.8, 0.90, 0.95$ and 1. For each r^{th} set of linked data a clerical sample of 300 links is selected. Each link in the clerical sample is assigned as being correct or incorrect. We summarise the performance of the ML estimator from section 3.2.2 and the naive method, which assumes there is no linkage error, by their 95% coverage rates and their Mean Squared Error (MSE). The coverage rates are based on the standard errors calculated from the Bootstrap described in section 3.3 with $R = 40$ replicates. The MSE of $\tilde{\beta}$ is calculated by

$$\text{MSE}(\tilde{\beta}) = \frac{1}{400} \sum_{r=1}^{400} (\tilde{\beta}_r - \beta)(\tilde{\beta}_r - \beta)'$$

where $\tilde{\beta}_r$ is the ML estimate of β from $\mathbf{d}^*(r)$.

Table 5 shows that the naive approach has poor coverage rates, due to its significant bias in the presence of linkage error, and consequently a relatively high MSE. The coverage rates for ML-Method 1 are very close to their nominal levels. The results show that, as the percentage of correct links reduces from 100% to 80%, the MSE of ML increases by a factor of about 3 for β_0 and β_1 . (The coverage rates and MSE of ML Method 1 and 2 were very similar so only the former are reported).

Table 5

Mean squared error and coverage rates for linked simulated data, where correct linkage occurs with probability, p

		Mean Squared Error				95% Coverage Rates		
		0.8	0.9	0.95	1	0.8	0.9	0.95
Naive	β_0	0.024	0.010	0.0056	0.0043*	0.35	0.80	0.93
	β_1	0.11	0.038	0.016	0.011*	0.05	0.62	0.88
ML-Method 1	β_0	0.013	0.0078	0.0055	0.0043*	93.0	94.25	93.5
	β_1	0.031	0.018	0.013	0.011*	96.0	94.5	96.25

*when $p=1$ the naive and ML estimators are the same by definition.

6. Discussion

Data linkage is an appropriate technique when data sets must be joined to enhance dimensions such as time and breadth or depth of detail. Data linkage is increasingly being used by statistical organisations around the world. It is well-known that errors can arise when linking files, for example when applying probabilistic linking methods. However, there has been little work reported in the literature about how to make valid inferences in the presence of such errors.

This paper provides methodological and practical advice to support analysts in this area.

In general, naively treating a linked file as if it were perfectly linked will lead to biased estimates. The analyst should only use the naive approach when both the number of unlinked records, defined as records that could be correctly linked but were not linked at all, and the number of incorrect links are negligible. This paper has presented a maximum likelihood approach to making valid inferences in the presence of both sources of error. The approach uses the well-known EM algorithm and is easy to apply in practice. The method can be applied when one of the files is not necessarily a subset of the other and when the linkage involves multiple passes. These situations often arise in practice, including many recent examples in the Australian Bureau of Statistics. The empirical study shows that the ML approach makes significant and meaningful improvements to the estimates from the linked data.

In the special case where File X is obtained by taking a random sample from File Y, the estimation procedure described is not 'full' maximum likelihood. This is because it does not use the fact that population totals for File Y are known. While inference using the method described here are still valid in this case, it could perhaps be made more efficient (see Scott and Wild 1997).

Acknowledgements

The authors would like to thank Raymond Chambers and two reviewers from Survey Methodology for their contributions to this paper.

References

- Australian Bureau of Statistics (2008). Census Data Enhancement - Indigenous Mortality Quality Study, 2006-07. Information Paper catalogue no. 4723.0.
- Bishop, G. (2009). Assessing the Likely Quality of the Statistical Longitudinal Census Dataset. Methodology Research Papers, catalogue no. 1351.0.55.026, Australian Bureau of Statistics, Canberra.
- Chambers, R., Chipperfield, J.O., Davis, W. and Kovačević, M. (2009). Regression Inference Based on Estimating Equations and Probability-Linked Data. Submitted for publication.
- Chambers, R.L., and Skinner, C.J. (2003). *Analysis of Survey Data*. New York: John Wiley & Sons, Inc.
- Chambers, R. (2008). Regression analysis of probability-linked data. *Statisphere*, Volume 4, <http://www.statisphere.govt.nz/official-statistics-research/series/vol-4.htm>.
- Christen, P., and Churches, T. (2005). Febri - Freely extensible biomedical record linkage. Release 0.3.1, viewed 17 November 2008, <http://cs.anu.edu.au/~Peter.Christen/Febri/febri-0.3/febri.doc-0.3/contents.html>.
- Conn, L., and Bishop, G. (2006). Exploring Methods for Creating a Longitudinal Census Dataset. Methodology Advisory Committee Papers, catalogue no. 1352.0.55.076, Australian Bureau of Statistics, Canberra.
- Fair, M. (2004). Generalized record linkage system-Statistics Canada's record linkage software. *Austrian Journal of Statistics*, 33(1 and 2), 37-53.
- Fellegi, I.P., and Sunter, A.B. (1969). A theory for record linkage. *Journal of the American Statistical Association*, 64, 1183-1210.
- Fuller, W.A. (1987). *Measurement Error Models*. New York: John Wiley & Sons, Inc.
- Hausman, J.A., Abrevaya, J. and Scott-Morton, F.M. (1998). Misclassification of the dependent variable in a discrete-response setting. *Journal of Econometrics*, 87, 239-269.
- Herzog, T.N., Scheuren, F.J. and Winkler, W.E. (2007). *Data Quality and Record Linkage Techniques*. New York: Springer.
- Holman, C.D.J., Bass, A.J., Rouse, I.L. and Hobbs, M.S.T. (1999). Population-based linkage of health records in Western Australia: Development of a health services research linked database. *Australian and New Zealand Journal of Public Health*, 23(5), 453-459.
- Lahiri, P., and Larsen, M.D. (2005). Regression analysis with linked data. *Journal of the American Statistical Association*, 100, 222-230.
- National Center for Health Statistics (2009). Linkages between Survey Data from the National Center for Health Statistics and Program Data from the Social Security Administration. Methodology Report, http://www.cdc.gov/nchs/data/datalinkage/ssa_methods_report_2009.pdf.
- Rubin, D.B., and Little, R.J.A. (2003). *Statistical analysis of missing data*, 2nd Edition. New York: John Wiley & Sons, Inc.
- Scheuren, F., and Winkler, W.E. (1993). Regression analysis of data files that are computer matched. *Survey Methodology*, 19, 39-58.
- Scott, A.J., and Wild, C.J. (1997). Fitting regression models to case-control data by maximum likelihood. *Biometrika*, 84, 57-71.
- Solon, R., and Bishop, G. (2009). A Linkage Method for the Formation of the Statistical Longitudinal Census Dataset. Methodology Research Papers, catalogue no. 1351.0.55.025, Australian Bureau of Statistics, Canberra.
- Winkler, W.E. (2001). Record Linkage Software and Methods for Merging Administrative Lists. Statistical Research Report Series, No. RR2001/03, Bureau of the Census.

Winkler, W.E. (2005). Approximate String Comparator Search Strategies for Very Large Administrative Lists. Statistical Research Report Series, no. RRS2005/02, Bureau of the Census.

Wright, J., Bishop, G. and Ayre, T. (2009). Assessing the Quality of Linking Migrant Settlement Records to Census Data. Methodology Research Papers, catalogue no. 1351.0.55.027, Australian Bureau of Statistics, Canberra.

Hierarchical Bayes small area estimation under a spatial model with application to health survey data

Yong You and Qian M. Zhou¹

Abstract

In this paper we study small area estimation using area level models. We first consider the Fay-Herriot model (Fay and Herriot 1979) for the case of smoothed known sampling variances and the You-Chapman model (You and Chapman 2006) for the case of sampling variance modeling. Then we consider hierarchical Bayes (HB) spatial models that extend the Fay-Herriot and You-Chapman models by capturing both the geographically unstructured heterogeneity and spatial correlation effects among areas for local smoothing. The proposed models are implemented using the Gibbs sampling method for fully Bayesian inference. We apply the proposed models to the analysis of health survey data and make comparisons among the HB model-based estimates and direct design-based estimates. Our results have shown that the HB model-based estimates perform much better than the direct estimates. In addition, the proposed area level spatial models achieve smaller CVs than the Fay-Herriot and You-Chapman models, particularly for the areas with three or more neighbouring areas. Bayesian model comparison and model fit analysis are also presented.

Key Words: Area level model; Bayesian model comparison; Disease rate; Gibbs sampling; Hierarchical spatial model; Posterior predictive model checking; Sampling variance.

1. Introduction

Model-based small area estimation methods have been widely used in practice due to the increasing demand for precise estimates for local regions and various small areas. In general sample surveys are designed to provide reliable estimates for large regions or aggregates of small areas such as the whole nation and provinces. Direct survey estimates, based only on the area specific sample data, usually provide reliable estimates of the parameter of interest for those large areas. For small areas, particularly some small geographical areas or specific small domains, direct estimates are likely to yield large standard errors because of the small sample sizes in those small areas. Therefore in making inference for small areas, it is necessary to borrow strength from related areas to form indirect estimates that increase the effective sample size and thus increase the precision of estimates. It is now generally accepted that the indirect estimates should be based on explicit models that provide links to related areas through the use of supplementary data such as census counts or administrative records; see, for example, Rao (2003) and Jiang and Lahiri (2006) for more discussion on model-based small area methods. The model-based estimates are obtained to improve the direct design-based estimates in terms of precision and reliability, *i.e.*, smaller coefficients of variation (CVs). There are two broad classifications for small area models: area level models and unit level models. Area level models are based on area direct survey estimates and unit level models are based on individual observations in small areas. In this paper we focus on area level models

that borrow strength across regions to improve the direct survey estimates.

Among the area level models, the Fay-Herriot model (Fay and Herriot 1979) is a basic and widely used area level model in practice to obtain reliable model-based estimates for small areas. The Fay-Herriot model basically has two components, namely, a sampling model for the direct estimates and a linking model for the parameters of interest. The sampling model involves the direct survey estimate and the corresponding sampling variance. The Fay-Herriot model assumes that the sampling variance is known in the model. Typically a smoothed estimator of the sampling variance is obtained and then treated as known in the model. Wang and Fuller (2003) and You and Chapman (2006) considered the situation where the sampling variances are unknown and modeled separately by direct estimators. In this paper we will consider both the smoothing and modeling methods for the sampling variances in the sampling model.

The linking model relates the parameter of interest to a regression model with area-specific random effects. In the Fay-Herriot model, the area random effects are usually assumed to be independent and identically distributed (*iid*) normal random variables to capture geographically unstructured variations among areas. However, in some small area applications, particularly in public health estimation problems, geographical variation of a disease is a subject of interest, and estimation of overall spatial pattern of risk and borrowing strength across regions to reduce variances of final estimates are both important. Thus, it may be more reasonable to construct spatial models on the area-specific

1. Yong You, Statistical Research and Innovation Division, Statistics Canada. E-mail: yongyou@statcan.gc.ca; Qian M. Zhou, Department of Biostatistics, Harvard University.

random effects to capture the spatial dependence among them. The spatial models are generally used in health related small area estimation, and various spatial models have been proposed for small area estimation (*e.g.*, Cressie 1990; Ghosh, Natarajan, Stroud and Carling 1998; Maiti 1998; Ghosh, Natarajan, Walter and Kim 1999; He and Sun 2000; Moura and Migon 2002; Singh, Shukla and Kundu 2005; Souza, Moura and Migon 2009). Best, Richardson and Thomson (2005) provided a comprehensive review on spatial models for disease mapping. Rao (2003) also discussed several spatial small area models.

The objective of this paper is to consider spatial correlation small area models and illustrate the usefulness of these models through an application to health survey data. The paper is organized as follows. In section 2, we first study area level models including the Fay-Herriot model and spatial correlation linking models. Then in section 3 we propose hierarchical Bayes (HB) small area models with spatial correlation and obtain HB inference for small area parameters through the Gibbs sampling method. In section 4, we apply the proposed models to the analysis of small area data from the Canadian Community Health Survey. We compare the performance of the model-based estimates with the direct design-based estimates, and moreover, we compare the proposed models with the Fay-Herriot model and the You-Chapman model (You and Chapman 2006) to investigate the effects of incorporating spatial structure on the area-specific random effects. Bayesian model comparison and model fit analysis are also provided. Finally in section 5, we offer some concluding remarks.

2. Small area models and inference

2.1 Fay-Herriot model

Let θ_i denote the parameter of interest for the i^{th} area, where $i = 1, \dots, m$, and m is the total number of areas. The Fay-Herriot model assumes that the θ_i 's are related to area specific auxiliary data $\mathbf{x}_i = (x_{i1}, \dots, x_{ip})'$ through a linear regression model as follows:

$$\theta_i = \mathbf{x}_i' \beta + v_i, \quad i = 1, \dots, m \quad (1)$$

where $\beta = (\beta_1, \dots, \beta_p)'$ is the $p \times 1$ vector of regression coefficients, and the v_i 's are area-specific random effects assumed to be *iid* with $E(v_i) = 0$ and $\text{Var}(v_i) = \sigma_v^2$. The assumption of normality may also be included. This model is referred to as a linking model for θ_i . The Fay-Herriot model also assumes that a direct survey estimator y_i , which is usually design-unbiased for the parameter of interest θ_i , is available whenever the area sample size $n_i > 1$. It is customary to assume that

$$y_i = \theta_i + e_i, \quad i = 1, \dots, m \quad (2)$$

where e_i 's are the sampling errors associated with the direct estimator y_i . We also assume that the e_i 's are independent normal random variables with mean $E(e_i | \theta_i) = 0$ and sampling variance $\text{Var}(e_i | \theta_i) = \sigma_i^2$. The model (2) is referred to as a sampling model for the direct survey estimator y_i . Combining these two components (1) and (2) leads to a linear mixed effects model (the Fay-Herriot model) as

$$y_i = \mathbf{x}_i' \beta + v_i + e_i, \quad i = 1, \dots, m. \quad (3)$$

In the basic Fay-Herriot model (3), the sampling variances σ_i^2 are usually assumed as known, which is a very strong assumption. Generally, we can use direct sampling variance estimates from the survey data, however, these direct estimates are unstable if sample sizes are small. Therefore, in practice, a smoothed estimator of σ_i^2 is used in the model and treated as known. A generalized variance function is usually applied in practice to obtain a smoothed estimator for the sampling variance, *e.g.*, Dick (1995). In recent years, a method of smoothing design effects has been developed and used in practice to obtain smoothed variance estimators (*e.g.*, Singh, Folsom and Vaish 2005; You 2008a; Liu, Lahiri and Kalton 2008). In particular, You (2008a) applied an equal design effects modeling approach to obtain smooth estimates of sampling variances. The design effect for the i^{th} area may be approximately written as

$$\text{deff}_i = \frac{s_i^2}{s_{ri}^2}, \quad \text{for } i = 1, \dots, m,$$

where s_i^2 is the unbiased direct estimate of sampling variance based on the complex sampling design, and s_{ri}^2 is the estimate of sampling variance based on the assumption of simple random sampling design. For each area, based on the assumption of a common design effect, a smoothed factor deff can be obtained by $\text{deff} = \sum_{i=1}^m \text{deff}_i / m$. Then a smoothed sampling variance estimate $\tilde{\sigma}_i^2$ can be obtained as $\tilde{\sigma}_i^2 = s_{ri}^2 \cdot \text{deff}$.

Instead of plugging in the smoothed estimates of sampling variances in the model, alternatively we can model the sampling variance directly. In the papers by Wang and Fuller (2003) and You and Chapman (2006), they assume the sampling variance σ_i^2 unknown and estimate σ_i^2 by an unbiased direct estimator s_i^2 , which is independent of the direct survey estimator y_i . They also assume that $d_i s_i^2 \sim \sigma_i^2 \chi_{d_i}^2$, where $d_i = n_i - 1$, and n_i is the sample size for the i^{th} area. You and Chapman (2006) considered the full HB approach with the Gibbs sampling method which automatically takes into account the extra uncertainty associated with the estimation of σ_i^2 . In this paper, we consider both the smoothing and modeling approaches for the sampling variances.

2.2 Spatial models

To incorporate spatially correlated random effects in the linking model, a simple and obvious way is to add a spatial random effect u_i in the independent linking model (1) as follows:

$$\theta_i = \mathbf{x}_i' \beta + v_i + u_i, \quad (4)$$

where u_i 's follow the well known intrinsic conditional autoregressive model given as

$$u_i | u_{-i} \sim N \left(\frac{\sum_{j \neq i} w_{ij} u_j}{\sum_{j \neq i} w_{ij}}, \frac{\sigma_u^2}{\sum_{j \neq i} w_{ij}} \right), \quad (5)$$

where u_{-i} denotes the values of spatial random effects u_j 's in all other areas with $j \neq i$, weights w_{ij} are fixed constants, and σ_u^2 is a unknown variance component. In practice, a common choice of w_{ij} is to let $w_{ij} = 0$ unless areas i and j are neighboring areas (*i.e.*, share a common boundary), in which case $w_{ij} = 1$. The model (4) is proposed by Besag, York and Mollie (1991) to separate spatial effects from overall heterogeneity in the areas. In model (4), independent random effects v_i capture geographically unstructured heterogeneity among areas, and spatial random effects u_i capture spatial dependence between areas. In this way, the degree of overall spatial dependence can be expressed based on the proportion of the total variation in $v_i + u_i$ captured by each component.

In practice, it is often unclear how to choose between an unstructured model (*e.g.*, the basic linking model) given by (1) and a purely spatially structured model (*e.g.*, intrinsic autoregressive model) given by (5). For model (4), posterior inference about the spatial dependence is based on the proportion of the total variation in the sum of $v_i + u_i$ captured by each component. However, although the univariate conditional distributions of the spatial component (5) are well defined, the corresponding joint distribution is improper (with undefined mean and infinite variance). Moreover, the model (4) has a potential identifiability problem where only the sum of the random effects $v_i + u_i$ is well identified by the data; see, for example, Best *et al.* (2005), for a more detailed discussion.

Alternatively, we can consider another spatial parameterization studied by Leroux, Lei, and Breslow (1999) and MacNab (2003), which avoids the identifiability problem encountered with the model (4). Let $\theta_i = \mathbf{x}_i' \beta + b_i$, and $\mathbf{b} = (b_1, \dots, b_m)'$. Following Leroux *et al.* (1999) and MacNab (2003), we place the following conditional autoregressive (CAR) model on the area specific spatial effects $\mathbf{b} = (b_1, \dots, b_m)'$:

$$\mathbf{b} \sim \text{MVN}(\mathbf{0}, \Sigma(\sigma_b^2, \lambda)) \quad (6)$$

$$\Sigma(\sigma_b^2, \lambda) = \sigma_b^2 \mathbf{D}^{-1}, \quad \mathbf{D} = \lambda \mathbf{R} + (1 - \lambda) \mathbf{I} \quad (7)$$

where σ_b^2 is a spatial dispersion parameter and λ is a spatial autocorrelation parameter, $0 \leq \lambda \leq 1$; \mathbf{I} is an identity matrix of dimension m ; \mathbf{R} , commonly known as the neighbourhood matrix, has i^{th} diagonal element equal to the number of neighbors of the area i , and the off-diagonal elements in each row equal to -1 if the corresponding areas are neighbors and 0 otherwise. The CAR model (6) - (7) corresponds to the following conditional distribution of b_i :

$$b_i | b_{-i} \sim N \left(\frac{\lambda}{1 - \lambda + \lambda w_{i+}} \sum_{j \neq i} w_{ij} b_j, \frac{\sigma_b^2}{1 - \lambda + \lambda w_{i+}} \right),$$

where $w_{i+} = \sum_{j \neq i} w_{ij}$. The CAR model (6) - (7) becomes the intrinsic autoregressive model (5) if $\lambda = 1$. On the other hand, if $\lambda = 0$, the CAR model (6) - (7) reduces to the independent linking model (1) which assumes independence on the area-specific random effects v_i . It is necessary to point out that the conditional mean and variances of $b_i | b_{-i}$ are weighted sums of the corresponding overall smoothing moments from the basic linking model (1) and local smoothing moments from the intrinsic autoregressive model:

$$\begin{aligned} E(b_i | b_{-i}) &= \frac{1 - \lambda}{1 - \lambda + \lambda w_{i+}} \times 0 \\ &\quad + \frac{\lambda w_{i+}}{1 - \lambda + \lambda w_{i+}} \left(\sum_{j \neq i} w_{ij} b_j / w_{i+} \right) \end{aligned}$$

$$\begin{aligned} \text{Var}(b_i | b_{-i}) &= \frac{1 - \lambda}{1 - \lambda + \lambda w_{i+}} \times \sigma_b^2 \\ &\quad + \frac{\lambda w_{i+}}{1 - \lambda + \lambda w_{i+}} (\sigma_b^2 / w_{i+}). \end{aligned}$$

Thus model (6)-(7) is a balance between the independent linking model (1) and the intrinsic CAR model (5). The spatial correlation parameter λ measures the extent of the spatial effects for local smoothing of the neighbouring areas. The modeling structure (6) captures both the unstructured heterogeneity among areas and the spatial correlation effects of the neighbouring area.

2.3 Hierarchical Bayes models and inference

In order to estimate θ_i , the parameter of interest, we apply a hierarchical Bayes (HB) approach using the Gibbs sampling method. Compared to other approaches such as EBLUP and empirical Bayes (EB), HB approach is straightforward and the inference for θ_i are exact unlike the EB or EBLUP. Moreover, the HB approach can deal with complex small area models using the Monte Carlo Markov Chain

(MCMC) method, which overcomes the computational difficulties of multi-dimensional integrations of posterior quantities to a large extent.

Let $\mathbf{y} = (y_1, \dots, y_m)'$, $\boldsymbol{\theta} = (\theta_1, \dots, \theta_m)'$, and $\mathbf{X} = (\mathbf{x}_1, \dots, \mathbf{x}_m)'$. We first construct two HB models without and with spatial structure under the assumption that the sampling variance σ_i^2 are assumed known and replaced by the smoothed estimate $\tilde{\sigma}_i^2$.

Model 1: Fay-Herriot model, denoted as FHM (Fay and Herriot 1979; Rao 2003).

- $y_i | \theta_i \sim N(\theta_i, \sigma_i^2 = \tilde{\sigma}_i^2)$, for $i = 1, \dots, m$;
- $\theta_i | \beta, \sigma_v^2 \sim N(\mathbf{x}_i' \beta, \sigma_v^2)$, for $i = 1, \dots, m$;
- Priors for the parameters (β, σ_v^2) : $\pi(\beta) \propto 1$; $\pi(\sigma_v^2) \sim \text{IG}(a_0, b_0)$, where a_0, b_0 are chosen to be very small known constants to reflect vague knowledge on σ_v^2 . N stands for the normal distribution and IG for the inverse gamma distribution.

Model 2: Proposed area level CAR model, as an extension of the Fay-Herriot model, denoted as CAR-FHM.

- $\mathbf{y} | \boldsymbol{\theta} \sim \text{MVN}(\boldsymbol{\theta}, \mathbf{E})$, where \mathbf{E} is a diagonal matrix with the i^{th} diagonal element $\sigma_i^2 = \tilde{\sigma}_i^2$;
- $\boldsymbol{\theta} | \beta, \sigma_v^2 \sim \text{MVN}(\mathbf{X}\beta, \sigma_v^2 \mathbf{D}^{-1})$, where $\mathbf{D} = \lambda \mathbf{R} + (1 - \lambda) \mathbf{I}$, with \mathbf{I} , an identity matrix of dimension m , and \mathbf{R} , the neighbourhood matrix;
- Priors for the parameters $(\beta, \lambda, \sigma_v^2)$: $\pi(\beta) \propto 1$; $\pi(\lambda) \sim \text{Uniform}(0, 1)$, where $0 \leq \lambda \leq 1$; $\pi(\sigma_v^2) \sim \text{IG}(a_0, b_0)$, where a_0, b_0 are chosen to be very small known constants. MVN stands for the multivariate normal distribution.

Note that the proposed model CAR-FHM reduces to FHM when the spatial autocorrelation parameter $\lambda = 0$.

We also consider two HB models with the sampling variance σ_i^2 unknown and modeled by the direct unbiased estimator s_i^2 .

Model 3: You-Chapman Model, denoted as YCM (You and Chapman 2006).

- $y_i | \theta_i, \sigma_i^2 \sim N(\theta_i, \sigma_i^2)$, for $i = 1, \dots, m$;
- $d_i s_i^2 | \sigma_i^2 \sim \sigma_i^2 \chi_{d_i}^2$ where $d_i = n_i - 1$, for $i = 1, \dots, m$;
- $\theta_i | \beta, \sigma_v^2 \sim N(\mathbf{x}_i' \beta, \sigma_v^2)$, for $i = 1, \dots, m$;
- Priors for unknown parameters $(\beta, \sigma_v^2, \sigma_i^2, i = 1, \dots, m)$: $\pi(\beta) \propto 1$; $\pi(\sigma_v^2) \sim \text{IG}(a_0, b_0)$, $\pi(\sigma_i^2) \sim \text{IG}(a_i, b_i)$ for $i = 1, \dots, m$, where a_i, b_i ($0 \leq i \leq m$) are chosen to be very small known constants to reflect vague knowledge on σ_i^2 and σ_v^2 .

Model 4: Proposed area level CAR model with unknown sampling variances, as an extension of You-Chapman model, denoted as CAR-YCM.

- $\mathbf{y} | \boldsymbol{\theta}, \sigma_1^2, \dots, \sigma_m^2 \sim \text{MVN}(\boldsymbol{\theta}, \mathbf{E})$, where matrix \mathbf{E} has diagonal elements σ_i^2 ;
- $d_i s_i^2 | \sigma_i^2 \sim \sigma_i^2 \chi_{d_i}^2$, where $d_i = n_i - 1$, for $i = 1, \dots, m$;
- $\boldsymbol{\theta} | \beta, \sigma_v^2 \sim \text{MVN}(\mathbf{X}\beta, \sigma_v^2 \mathbf{D}^{-1})$, where $\mathbf{D} = \lambda \mathbf{R} + (1 - \lambda) \mathbf{I}$;
- Priors for the parameters $(\beta, \lambda, \sigma_v^2, \sigma_i^2, i = 1, \dots, m)$: $\pi(\beta) \propto 1$; $\pi(\lambda) \sim \text{Uniform}(0, 1)$, where $0 \leq \lambda \leq 1$; $\pi(\sigma_v^2) \sim \text{IG}(a_0, b_0)$; $\pi(\sigma_i^2) \sim \text{IG}(a_i, b_i)$ for $i = 1, \dots, m$, where a_i, b_i ($0 \leq i \leq m$) are chosen to be very small known constants.

Again, note that the proposed model CAR-YCM reduces to the You-Chapman model when $\lambda = 0$. For both models 3 and 4 there is an implicit assumption that the area-specific sample size $n_i \geq 2$. If flat priors are used for σ_i^2 , we should have $n_i \geq 4$ to ensure proper posteriors (You and Chapman 2006).

We apply the Gibbs sampling method to estimate the posterior mean $E(\theta_i | \mathbf{y})$ and the corresponding posterior variance $\text{Var}(\theta_i | \mathbf{y})$. The required full conditional distributions of parameters under different models are given in Appendix A. For the Fay-Herriot model and the You-Chapman model, all the full conditional distributions have closed forms and drawing samples from these distributions is straightforward. For the proposed two area level spatial models CAR-FHM and CAR-YCM, the conditional distribution of the spatial correlation parameter λ does not have a closed form. We use the Metropolis-Hastings algorithm within the Gibbs sampler (Chip and Greenberg 1995) to update λ . Under the model CAR-FHM, the full conditional distribution of λ in the Gibbs sampler can be written as

$$[\lambda | \boldsymbol{\theta}, \beta, \sigma_v^2] \propto h(\lambda) f(\lambda)$$

where $f(\lambda)$ is a density function of the uniform distribution, $\text{Uniform}(0, 1)$, given as

$$f(\lambda) \propto 1, \text{ where } 0 \leq \lambda \leq 1$$

and $h(\lambda)$ is a function given by

$$h(\lambda) \propto \left| [\lambda \mathbf{R} + (1 - \lambda) \mathbf{I}]^{-1} \right|^{-1/2} \times \exp \left\{ -\frac{1}{2\sigma_v^2} (\boldsymbol{\theta} - \mathbf{X}\beta)' [\lambda \mathbf{R} + (1 - \lambda) \mathbf{I}] (\boldsymbol{\theta} - \mathbf{X}\beta) \right\}.$$

We use $f(\lambda)$ as the “candidate” generating density function in the Metropolis-Hastings updating step. To update λ from the current values of $(\theta^{(k)}, \beta^{(k)}, \sigma_v^{2(k)})$, we proceed as follows:

1. Draw λ^* from a uniform distribution;
2. Compute the acceptance probability $\alpha(\lambda^*, \lambda^{(k)}) = \min\{h(\lambda^*)/h(\lambda^{(k)}), 1\}$;
3. Generate u from a uniform distribution, if $u < \alpha(\lambda^*, \lambda^{(k)})$, then the candidate value λ^* is accepted, i.e., $\lambda^{(k+1)} = \lambda^*$; otherwise λ^* is rejected, and set $\lambda^{(k+1)} = \lambda^{(k)}$.

For the model CAR-YCM, a similar procedure can be applied when drawing samples from the conditional distribution of λ .

3. Data analysis

3.1 Data description and implementation

The Canadian Community Health Survey (CCHS) is a federal survey conducted by Statistics Canada. The primary objective of CCHS is to provide timely and reliable estimates of health determinants, health status and health system utilization across Canada. It is a cross-sectional survey which operates on a two-year collection cycle. The first year of the survey cycle “x.1” targets individuals aged 12 or older who are living in private dwellings, and it is a general population health survey with a large sample (130,000 persons) designed to provide reliable estimates at the health region, provincial and national levels. The second year of the survey cycle “x.2” has a smaller sample (30,000 persons) allocated based on provincial sample buy-ins and is designed to provide provincial and national level results on specific focused health topics. Although national and provincial estimates are very important, there is an increasing demand for health data at lower levels of geography voiced by a number of provinces including British Columbia (BC), Prince Edward Island (PEI), Quebec and others. Cycle “x.1” of the CCHS collected data corresponds to 136 health regions in the 10 provinces and three territories. It primarily used two sampling frames. The first one, used as the primary frame, was based on the area frame designed for the Canadian Labour Force Survey, and within the area frame, a multistage stratified cluster design was used to sample dwellings. The second frame consists of a list of telephone numbers. Random digit dialing methodology is used in some of the health regions for cost reasons. More details of the design are provided in Béland (2002). In this paper, we use a small data set from Cycle 1.1 as an example to demonstrate the analysis. We are interested in estimating the disease rate for local health regions within

provinces. In particular, we apply the four models discussed in section 2 to estimate the asthma rate for 20 health regions in the province of BC using the data from Cycle 1.1. Figure 1 shows the map of the 20 health regions in the province of British Columbia. We use this map to define the neighbourhood correlation matrix used in the spatial models. Appendix B gives the list of health regions and related spatial structures.

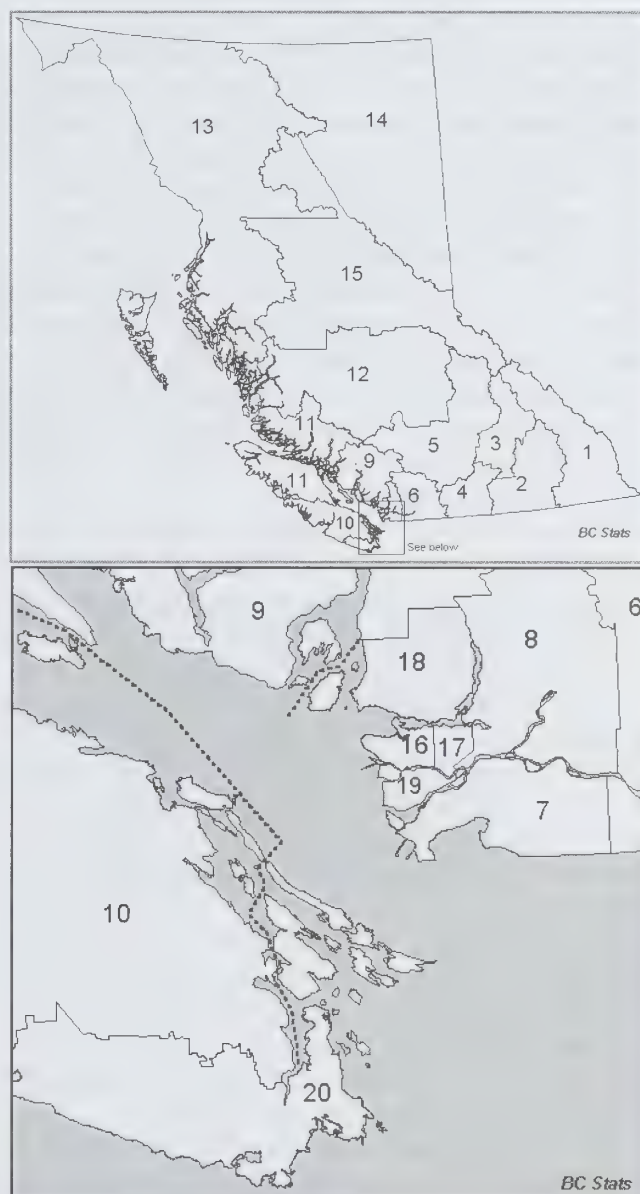


Figure 1 Map of 20 health regions in the province of British Columbia

Let θ_i denote the true asthma rate for the i^{th} health region in BC, $i = 1, \dots, 20$. From the survey data of Cycle 1.1, we obtained the direct survey estimate y_i of θ_i as the ratio of number of people having asthma (direct survey

estimate) divided by the corresponding population size (known constant). We have also included six area level auxiliary variables used in the model, and these six variables are total population size, number of persons who have asthma as one of the symptoms of the chronic disease, number of persons who have asthma as the main symptom of the chronic disease, number of persons who have diabetes as one of the symptoms of the chronic disease, number of persons who have diabetes as the main symptom of the chronic disease, and number of visits to hospitals. Note that in the literature related to disease mapping (e.g., Mollié 1996; Maiti 1998; MacNab 2003), a Poisson or Binomial distribution is usually assumed in the sampling model for the direct estimate y_i . However, in small area estimation, the direct estimate y_i is obtained based on the complex sampling design used in the survey. Thus, it is a customary approach to assume a normal sampling model on the direct estimates y_i ; see, for example, Datta, Lahiri, Maiti and Lu (1999), Rao (2003), Mohadjer, Rao, Liu, Krenzke and Van de Kerckhove (2007), and You (2008a). Note that we have only considered one kind of disease rate data from one province in our study and used this example as illustration of the proposed model and evaluate the effects of spatial modeling in small area models.

To implement the Gibbs sampling, we use $L = 5$ parallel runs each with a "burn-in" length of $B = 2,000$ and Gibbs sampling size of $G = 5,000$. For the proposed models CAR-FHM and CAR-YCM, in order to reduce the autocorrelation which results from the accept-rejection algorithm in the run, we take every 5th iteration after the "burn-in" period. Therefore, for models FHM and YCM, we have $n = 5,000$ samples for each run, and for models CAR-FHM and CAR-YCM, we have $n = 1,000$ samples for each run. Convergence of the Gibbs sampling is monitored for the small area parameters θ_i and other unknown parameters in the model using the potential scale reduction factor (Gelman and Rubin 1992; Gelman, Carlin, Stern and Rubin 2004, page 296-297). We have computed the reduction factors for all the monitored parameters in the model in the Gibbs sampling. These factor values are all very close to 1 (less than 1.05), which suggests that the desired convergence for these parameters is achieved by the Gibbs sampler.

We have used vague priors for the hyperparameters in the model as a common practice in HB small area estimation. In particular, the flat prior for regression parameter $\pi(\beta) \propto 1$ and proper inverse gamma priors for variance components are commonly used (e.g., Arora and Lahiri 1997; Ghosh *et al.* 1998; Datta *et al.* 1999; You and Rao 2000; Rao 2003, page 237; Souza *et al.* 2009). Following MacNab (2003), we have used the uniform prior $\pi(\lambda) \sim \text{Uniform}(0, 1)$ for the autocorrelation parameter. The uniform priors are also commonly used for the autocorrelation

parameters in spatial models (e.g., Maiti 1998; He and Sun 2000; Rao 2003, page 266). We also tried several different values for the inverse gamma priors. The HB estimates are quite stable and not sensitive to the choice of vague proper priors. More detailed discussion on sensitivity analysis can be found, for example, in You and Chapman (2006) for similar models.

3.2 Comparison of results

At first, we present the HB estimates of the asthma rate under models FHM and CAR-FHM in which the sampling variances σ_i^2 are assumed to be known. We used the smoothed estimate $\tilde{\sigma}_i^2$ obtained by the smoothing technique in You (2008a) as described in Section 2. Figure 2 displays the direct estimates and the HB model-based estimates under FHM and CAR-FHM for the 20 health regions in BC. The health regions appear in the x-coordinate ranked by the order of sample size with the smallest (Peace Liard) on the left and the largest (South Fraser Valley) on the right. Model 1 (FHM) and Model 2 (CAR-FHM) give similar point estimates, and both the model-based estimates lead to moderate smooth estimates compared to the direct estimates. Moreover, the direct estimates and two HB estimates of the disease rate are very close for some health regions with large sample sizes, but for some areas with smaller sample sizes, they differ to some extent. Similar results are obtained under Model 3 (YCM) and Model 4 (CAR-YCM).

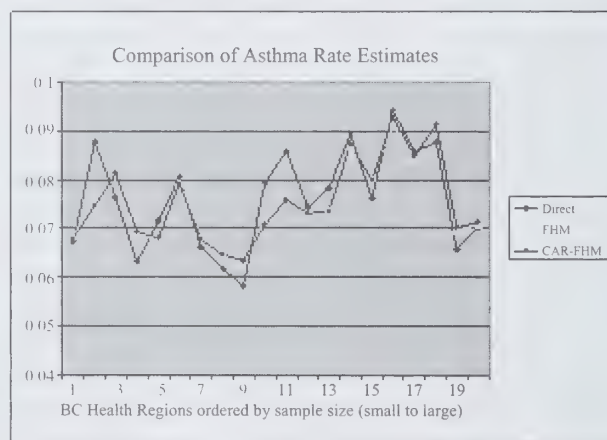


Figure 2 Direct and HB model-based estimates under models FHM and CAR-FHM

Figure 3 presents the CVs of the direct and two HB model-based estimates with the health regions ordered by the sample sizes from the smallest to the largest as in Figure 2. The CVs of HB estimates are obtained by dividing the squared root of the posterior variance by the posterior mean. As expected, the CVs of the direct estimates show a clear tendency of decrease as the sample size increases. However, the two model-based estimates give smoother CVs. Moreover, the two HB model-based estimates exhibit a great

improvement over the direct design-based estimates in terms of precision and reliability, that is, smaller CVs. Compared to the direct estimates, the average CV reduction of the HB estimates under FHM is about 22.7% ranging from 7.8% to 40.5%, and the average reduction of the CVs for the HB estimates under the proposed CAR-FHM is 27.8% ranging from 12.5% to 52.1%. Thus it is clear that the proposed spatial model CAR-FHM is superior to the Fay-Herriot model. We also obtained similar results for the models YCM and CAR-YCM when the sampling variance is modeled directly. The average CV reduction under YCM is 23.9%, whereas the average CV reduction is 29.0% under the proposed spatial model CAR-YCM. Details of the results including the point estimates and the corresponding CVs are presented in a table in Appendix C. In our example, the sample size at the health region level is relatively large. The model-based estimates have still shown great improvement over the direct survey estimates. Our results indicate that the presented small area models can be used to improve the direct survey estimates even when the sample size is relatively large. Note that Bayesian credible intervals for the small area parameters can be easily constructed using the MCMC output from the Gibbs sampler if required by practical users. This is an advantage of using the HB inference via MCMC sampling. However in this paper we only report the model-based point estimates and the corresponding CVs as our main purpose is to compare the model-based estimates with the direct estimates and to show the efficiency gain of the models. The gain in efficiency is clearly evident.

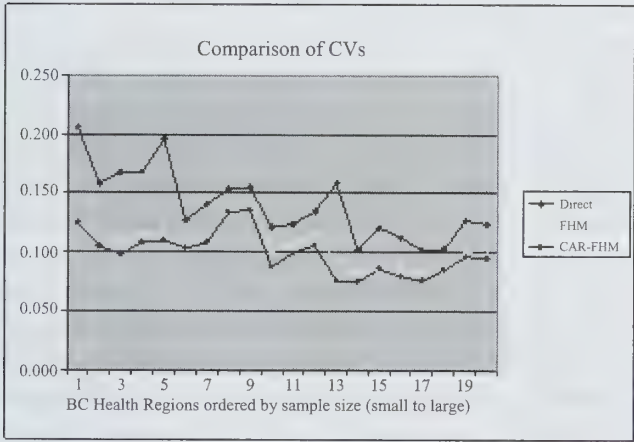


Figure 3 Direct and HB CVs under models FHM and CAR-FHM

In order to investigate the effects of incorporating the spatial structure in the model, we present the CVs of the direct and HB estimates by health regions sorted according to the number of neighbouring regions from the smallest (2 neighbours) to the largest (7 neighbours) in Figure 4. It shows that the HB estimates from the proposed model

CAR-FHM has smaller CVs than the estimates from the Fay-Herriot model. In addition, the improvement of CAR-FHM over the Fay-Herriot model is much more obvious in the regions with more neighbours, and these two models give very close CVs in the regions with less adjacent areas. Very similar results are also obtained for CAR-YCM over YCM. Table 1 gives the average reduction of the CVs across the health regions with the same number of neighbours. The results in Table 1 present the CV reduction of the proposed spatial models for both cases of known and unknown sampling variances. For example, for known σ_i^2 (smoothed $\tilde{\sigma}_i^2$), for areas with only 2 neighbours, the average CV reduction of model CAR-FHM over the Fay-Herriot model is only around 0.9%, whereas for areas with 7 neighbours, the average CV reduction for CAR-FHM over FHM is as high as around 20%. For the case of unknown σ_i^2 , similar results are obtained for CAR-YCM over YCM. The numerical results in Table 1 confirm the clear trend of increased CV reduction under the proposed spatial model over FHM or YCM as the number of neighbours increases. Thus, more neighbouring areas can provide more information in the spatial structure to improve the precision and reliability of the HB estimates.

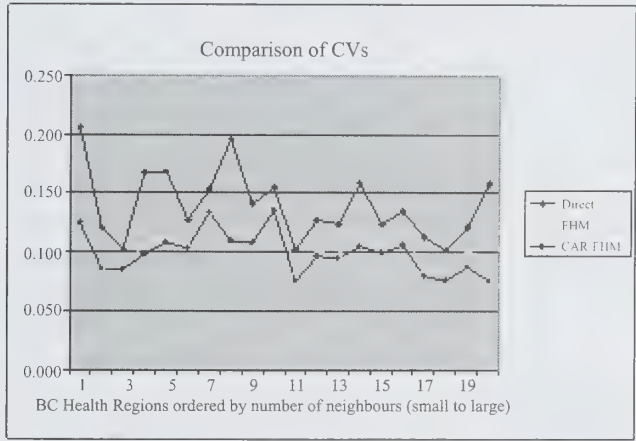


Figure 4 Direct and HB CVs under models FHM and CAR-FHM with the health regions sorted by the number of neighbours

Table 1
Comparison of average CV reduction

Number of neighbours	Average CV reduction	
	CAR-FHM over FHM	CAR-YCM over YCM
2	0.9%	1.8%
3	3.7%	3.5%
4	6.3%	6.0%
5	8.9%	8.7%
6	13.7%	11.0%
7	19.2%	20.7%

3.3 Bayesian model comparison

In this section, we compare the proposed models CAR-FHM with FHM and CAR-YCM with YCM, respectively. For hierarchical Bayes model comparison, the deviance information criterion (DIC) proposed by Spiegelhalter, Best, Carlin and van der Linde (2002) is commonly used in recent years to compare non-nested and mixed effects Bayesian models. The DIC is based on the deviance of the model $D(\theta)$, which is equal to minus twice the log-likelihood of the model, and the DIC is usually computed as $DIC = D(\hat{\theta}) + 2p_D$, where $D(\hat{\theta})$ is the deviance of the model evaluated at the posterior mean of the model parameters, which summarizes the goodness of fit of the model, and p_D is the effective number of parameters, which captures the complexity of the model. p_D is defined as $p_D = \bar{D}(\theta) - D(\hat{\theta})$, and $\bar{D}(\theta)$ is the posterior mean of the deviance of the model. Thus the DIC is defined as the summation of the goodness of fit of the model and the model complexity. Smaller values of DIC indicate a better model fit. Computation of DIC is relatively straightforward provided that the deviance $D(\theta)$ is available in closed form, and p_D may be calculated after the Gibbs sampling run by taking the sample mean of the simulated values of $D(\theta)$ minus the plug-in estimate of the deviance $D(\hat{\theta})$. For the four models presented in section 2, we computed the corresponding DIC values, as shown in Table 2. It is clear that the proposed spatial models CAR-FHM and CAR-YCM both have smaller DIC values than the non-spatial models FHM and YCM respectively, which indicates that the spatial models are better than the non-spatial models in our study. Both spatial models CAR-FHM and CAR-YCM perform well in this example. This result of model comparison is consistent with the estimation results presented in section 3.2.

Table 2
Comparison of DIC values for the four hierarchical models

Model	DIC value
FHM	27.1
CAR-FHM	24.6
YCM	26.8
CAR-YCM	24.5

3.4 Test of model fit

In order to check the overall model fit of the proposed models CAR-FHM and CAR-YCM, we use the method of posterior predictive distribution. Let y_{rep} denote the replicated observation under the model. The posterior predictive distribution of y_{rep} given the observed data y_{obs} is defined as $f(y_{rep} | y_{obs}) = \int f(y_{rep} | \theta) f(\theta | y_{obs}) d\theta$. In this approach, a test statistic $T(y, \theta)$ that depends on the data y and possibly the parameter θ can be defined and the

observed value $T(y_{obs}, \theta | y_{obs})$ compared to the posterior predictive distribution of $T(y_{rep}, \theta | y_{obs})$ with any significant difference indicates a model failure. Lack of fit of the data with respect to the posterior predictive distribution can be measured by the p -value of the test quantity (Meng 1994; Gelman, Meng and Stern 1996). The posterior predictive p -value is defined as $p = P(T(y_{rep}, \theta) \geq T(y_{obs}, \theta) | y_{obs})$. If the given model adequately fits the observed data, then $T(y_{obs}, \theta | y_{obs})$ should be near the central part of the histogram of the $T(y_{rep}, \theta | y_{obs})$ values if y_{rep} is generated repeatedly from the posterior predictive distribution. Consequently, the posterior predictive p -value is expected to be near 0.5 if the model adequately fits the data. Extreme p -values (near 0 or 1) suggest poor fit. The posterior predictive p -value model checking has been criticized for being conservative due to the double use of the observed data; see, for example, Bayarri and Berger (2000). They proposed alternative model checking p -value measures, named the partial posterior predictive p -value and the conditional predictive p -value. However, their methods are more difficult to implement and interpret (Rao 2003; Sinharay and Stern, 2003). As noted in Sinharay and Stern (2003), the posterior predictive p -value is especially useful if we think of the current model as a plausible ending point with modifications to be made only if substantial lack of fit is found.

To carry out the posterior predictive model checking, we need to specify a test quantity $T(y, \theta)$. You (2008b) studied several test quantities in posterior predictive model checking for small area models through a simulation study and proposed a test quantity given as

$$T(y, \theta) = |\max(y_i) - \text{mean}(\theta_i)| - |\min(y_i) - \text{mean}(\theta_i)|.$$

It is shown in You (2008b) that the proposed test quantity $T(y, \theta)$ is sensitive to the choice of distribution of random effects and different mean functions under the Fay-Herriot model. A similar test quantity is also suggested in Gelman *et al.* (2004) for posterior predictive model checking. In our study, under the proposed model CAR-FHM, the estimated p -value is 0.472, and under model CAR-YCM, the estimated p -value is 0.453. Thus there is no indication of lack of model fit and both proposed spatial models fit the data quite well.

To access model fit at the individual observation level, we also computed the individual predictive probability values p_i^* as $p_i^* = P(y_{i(rep)} < y_{i(obs)} | y_{obs})$; see, for example, Gelfand (1996) and Daniels and Gatsonis (1999). These individual predictive probabilities provide information on the degree of consistent overestimation or underestimation of the observed data. For model CAR-FHM, the p_i^* ranges from 0.325 to 0.768 with a mean of 0.517 and a median of 0.496; for model CAR-YCM, the p_i^* ranges from 0.316 to

0.772 with a mean of 0.511 and a median of 0.497. Both models give very similar results and the mean and median values are all around 0.5. There is no indication of any consistent overestimation or underestimation of the proposed models. The overall p -values and individual predictive probabilities have shown that the proposed spatial small area models fit the data quite well.

3.5 Bias diagnostics

To evaluate any possible bias of the model-based estimates under the proposed models with respect to the direct survey estimates, following Brown, Chambers, Heady and Heasman (2001), we consider a simple method of regression analysis for the direct estimates and the HB model-based estimates. You (2008a) also used the regression analysis method for model bias diagnostics. If the model-based estimates are close to the true values of the small area disease rate, then the direct survey estimates, which are assumed to be unbiased for the true disease rates, should behave like random variables whose expected values correspond to the values of the model-based estimates. That means the model-based estimates should be unbiased predictors of the direct estimates. In terms of regression analysis, we basically fit the regression model $Y = \alpha + \beta X$ to the data and estimate the coefficients, and see how close the regression line is to $Y = X$. Let Y be the direct survey estimates and X be the model-based estimates. Under the proposed CAR-FHM, we obtain a regression line $Y = -0.0021(0.011) + 1.0365(0.1445)X$; under the proposed CAR-YCM, we obtain a regression line $Y = -0.0028(0.0108) + 1.0458(0.1427)X$. Thus both the regression lines show very little disparity from $Y = X$. We therefore conclude that the model-based estimates are consistent with the direct estimates with no extra possible bias induced by the proposed models. The results also provide an indication of no evidence of any bias due to possible model misspecification.

4. Conclusions

In this paper we have discussed two area level models, namely, the well-known Fay-Herriot model in which the sampling variance is assumed to be known, and the You-Chapman model in which the sampling variance is unknown and modeled separately by its direct estimator. In both the Fay-Herriot model and You-Chapman model, the area random effects are assumed to be *iid* normal random variables to capture unexplained area heterogeneity effects. After comparing various forms of Gaussian CAR models proposed in the literature (e.g., Best *et al.* 2005) for disease mapping to incorporate spatially correlated effects, we extended the independent area effects model to a spatial correlation model and combined it with the traditional small

area models. The proposed new small area spatial correlation models CAR-FHM and CAR-YCM include the small area sampling models and a spatial correlation linking model which captures both the unstructured heterogeneity among areas and the spatial correlation effects of the neighbouring areas. We don't need to specify the spatial autocorrelation parameter in the model, and this parameter will be estimated from the data.

In the data analysis we compared the proposed spatial models with the non-spatial effects models by applying the models to estimate the rates of asthma for 20 health regions in the province of British Columbia. Our results have shown that the model-based estimates achieve a great improvement over the direct estimates in terms of moderately smoothed point estimates and much smaller CVs. Particularly, the proposed models are superior to the Fay-Herriot model or You-Chapman model whether the sampling variances are assumed to be known or unknown. Moreover, note that the CV reduction of the proposed spatial models over the Fay-Herriot model or You-Chapman model is greater for the areas with more neighbours. Results of the Bayesian model comparison and model fit analysis are also in favor of the proposed small area spatial models.

In future work, the proposed small area spatial models can be extended to unmatched sampling and linking models (You and Rao 2002) with the sampling variance known or unknown. We plan to evaluate the estimation effects of different spatial models as well as the effects of spatial structures. For data analysis, we will produce model-based health status estimates based on the proposed models for health regions across Canada and evaluate the possibility of extending the model-based approach to lower level estimates such as age-sex domains within health regions. We also plan to consider the data cloning method (Lele, Dennis and Lutscher 2007; Lele, Nadeem and Schmuland 2010) for the spatial models. An advantage of data cloning method is that the results are independent of the choice of priors. But the computational burden could be considerably extensive.

Acknowledgements

We would like to thank one Associate Editor and one referee for their detailed comments and suggestions. Yong You's research work was supported by Statistics Canada Methodology Branch Research Block Fund. Qian M. Zhou's work was finished when she worked at Statistics Canada as a MITACS/NPCDS research internship student under the supervision of Yong You. Q.M. Zhou presented the proposed models and some results of the paper at the 2008 Statistics Society of Canada (SSC) annual meeting in Ottawa, and won the 2008 best student paper award of the SSC Survey Methods Section.

Appendix A

Full conditional distributions

A.1. Gibbs sampling full conditional distributions under Model 1: FHM.

- $[\theta_i | y_i, \beta, \sigma_v^2] \sim N[\gamma_i y_i + (1 - \gamma_i) \mathbf{x}_i' \beta, \tilde{\sigma}_i^2 \gamma_i]$, where $\gamma_i = \sigma_v^2 / (\sigma_v^2 + \tilde{\sigma}_i^2)$, for $i = 1, \dots, m$;
- $[\beta | \boldsymbol{\theta}, \sigma_v^2] \sim N \left[\left(\sum_{i=1}^m \mathbf{x}_i \mathbf{x}_i' \right)^{-1} \left(\sum_{i=1}^m \mathbf{x}_i \theta_i \right), \sigma_v^2 \left(\sum_{i=1}^m \mathbf{x}_i \mathbf{x}_i' \right)^{-1} \right]$;
- $[\sigma_v^2 | \boldsymbol{\theta}, \beta] \sim \text{IG} \left[a_0 + \frac{1}{2}m, b_0 + \frac{1}{2} \sum_{i=1}^m (\theta_i - \mathbf{x}_i' \beta)^2 \right]$.

A.2. Gibbs sampling full conditional distributions under Model 2: CAR-FHM.

- $[\boldsymbol{\theta} | \mathbf{y}, \beta, \lambda, \sigma_v^2] \sim \text{MVN}(\boldsymbol{\Lambda} \mathbf{y} + (\mathbf{I} - \boldsymbol{\Lambda}) \mathbf{X} \beta, \boldsymbol{\Lambda} \mathbf{E})$, where $\boldsymbol{\Lambda} = (\mathbf{E}^{-1} + \mathbf{D} / \sigma_v^2)^{-1} \mathbf{E}^{-1}$ with $\mathbf{E} = \text{diag} \{ \tilde{\sigma}_1^2, \dots, \tilde{\sigma}_m^2 \}$ and $\mathbf{D} = \lambda \mathbf{R} + (1 - \lambda) \mathbf{I}$;
- $[\beta | \boldsymbol{\theta}, \lambda, \sigma_v^2] \sim \text{MVN}[(\mathbf{X}' \mathbf{D} \mathbf{X})^{-1} \mathbf{X}' \mathbf{D} \boldsymbol{\theta}, \sigma_v^2 (\mathbf{X}' \mathbf{D} \mathbf{X})^{-1}]$;
- $[\lambda | \boldsymbol{\theta}, \beta, \sigma_v^2] \propto |\lambda \mathbf{R} + (1 - \lambda) \mathbf{I}|^{-1/2} \times \exp \left\{ -\frac{1}{2\sigma_v^2} (\boldsymbol{\theta} - \mathbf{X} \beta)' [\lambda \mathbf{R} + (1 - \lambda) \mathbf{I}] (\boldsymbol{\theta} - \mathbf{X} \beta) \right\}$;
- $[\sigma_v^2 | \boldsymbol{\theta}, \beta, \lambda] \sim \text{IG} \left[a_0 + \frac{m}{2}, b_0 + \frac{1}{2} (\boldsymbol{\theta} - \mathbf{X} \beta)' \mathbf{D} (\boldsymbol{\theta} - \mathbf{X} \beta) \right]$.

A.3. Gibbs sampling full conditional distributions under Model 3: YCM.

- $[\theta_i | y_i, \beta, \sigma_i^2, \sigma_v^2] \sim N[\gamma_i y_i + (1 - \gamma_i) \mathbf{x}_i' \beta, \sigma_i^2 \gamma_i]$, where $\gamma_i = \sigma_v^2 / (\sigma_v^2 + \sigma_i^2)$, for $i = 1, \dots, m$;
- $[\beta | \boldsymbol{\theta}, \sigma_v^2] \propto N \left[\left(\sum_{i=1}^m \mathbf{x}_i \mathbf{x}_i' \right)^{-1} \left(\sum_{i=1}^m \mathbf{x}_i \theta_i \right), \sigma_v^2 \left(\sum_{i=1}^m \mathbf{x}_i \mathbf{x}_i' \right)^{-1} \right]$;
- $[\sigma_i^2 | y_i, \theta_i] \sim \text{IG} \left(a_i + \frac{d_i + 1}{2}, b_i + \frac{(y_i - \theta_i)^2 + d_i s_i^2}{2} \right)$, where $d_i = n_i - 1$, for $i = 1, \dots, m$;
- $[\sigma_v^2 | \boldsymbol{\theta}, \beta] \sim \text{IG} \left[a_0 + \frac{1}{2}m, b_0 + \frac{1}{2} \sum_{i=1}^m (\theta_i - \mathbf{x}_i' \beta)^2 \right]$.

A.4. Gibbs sampling full conditional distributions under Model 4: CAR-YCM.

- $[\boldsymbol{\theta} | \mathbf{y}, \beta, \lambda, \sigma_v^2, \sigma_i^2] \sim \text{MVN}(\boldsymbol{\Lambda} \mathbf{y} + (\mathbf{I} - \boldsymbol{\Lambda}) \mathbf{X} \beta, \boldsymbol{\Lambda} \mathbf{E})$, where $\boldsymbol{\Lambda} = (\mathbf{E}^{-1} + \mathbf{D} / \sigma_v^2)^{-1} \mathbf{E}^{-1}$, and $\mathbf{E} = \text{diag} \{ \sigma_1^2, \dots, \sigma_m^2 \}$, $\mathbf{D} = \lambda \mathbf{R} + (1 - \lambda) \mathbf{I}$;
- $[\beta | \boldsymbol{\theta}, \lambda, \sigma_v^2] \sim \text{MVN}[(\mathbf{X}' \mathbf{D} \mathbf{X})^{-1} \mathbf{X}' \mathbf{D} \boldsymbol{\theta}, \sigma_v^2 (\mathbf{X}' \mathbf{D} \mathbf{X})^{-1}]$;
- $[\lambda | \boldsymbol{\theta}, \beta, \sigma_i^2] \propto |\lambda \mathbf{R} + (1 - \lambda) \mathbf{I}|^{-1/2} \times \exp \left\{ -\frac{1}{2\sigma_v^2} (\boldsymbol{\theta} - \mathbf{X} \beta)' [\lambda \mathbf{R} + (1 - \lambda) \mathbf{I}] (\boldsymbol{\theta} - \mathbf{X} \beta) \right\}$;
- $[\sigma_i^2 | y_i, \theta_i] \sim \text{IG} \left(a_i + \frac{d_i + 1}{2}, b_i + \frac{(y_i - \theta_i)^2 + d_i s_i^2}{2} \right)$, where $d_i = n_i - 1$, for $i = 1, \dots, m$;
- $[\sigma_v^2 | \boldsymbol{\theta}, \beta, \lambda] \sim \text{IG} \left[a_0 + \frac{m}{2}, b_0 + \frac{1}{2} (\boldsymbol{\theta} - \mathbf{X} \beta)' \mathbf{D} (\boldsymbol{\theta} - \mathbf{X} \beta) \right]$.

Appendix B

**List of 20 health regions in the province
of British Columbia with the corresponding sample sizes and spatial structures**

ID number	Health region name	Sample size	Number of neighbours	Neighbours
1	East Kootenay	645	3	2, 3, 15
2	West Kootenay-Boundary	705	3	1, 3, 4
3	North Okanagan	890	5	1, 2, 4, 5, 15
4	South Okanagan Similameen	1,063	4	2, 3, 5, 6
5	Thompson	982	7	3, 4, 6, 9, 11, 12, 15
6	Fraser Valley	1,125	5	4, 5, 7, 8, 9
7	South Fraser Valley	1,437	4	6, 8, 17, 19
8	Simon Fraser	1,165	5	6, 7, 9, 17, 18
9	Coast Garibaldi	623	5	5, 6, 8, 11, 18
10	Central Vancouver Island	1,077	2	11, 20
11	Upper Island/Central Coast	746	4	5, 9, 10, 12
12	Cariboo	673	4	5, 11, 13, 15
13	North West	650	3	12, 14, 15
14	Peace Liard	611	2	13, 15
15	Northern Interior	859	6	1, 3, 5, 12, 13, 14
16	Vancouver	1,285	4	17, 18, 19, 20
17	Burnaby	871	5	7, 8, 16, 18, 19
18	North Shore	842	4	8, 9, 16, 17
19	Richmond	828	3	7, 16, 17
20	Capital	1,225	2	10, 16

Note that Vancouver (#16) and Capital (#20) are not adjacent regions in the map since they are separated by the ocean. However, due to the intensive and close connection between these two regions, we define them as neighbours in our study for illustration purpose only.

Appendix C

Direct and model-based point estimates and CVs

Area ID	Direct Est.	Comparison of point estimates			
		FHM	CAR-FHM	YCM	CAR-YCM
1	0.0765	0.0793	0.0812	0.0795	0.0812
2	0.0804	0.0795	0.0793	0.0797	0.0794
3	0.0745	0.0726	0.0731	0.0725	0.0729
4	0.0893	0.0868	0.0874	0.0867	0.0873
5	0.0782	0.0739	0.0736	0.0729	0.0731
6	0.0943	0.0914	0.0927	0.0918	0.0928
7	0.0702	0.0707	0.0712	0.0711	0.0717
8	0.0858	0.0845	0.0848	0.0844	0.0849
9	0.0877	0.0763	0.0745	0.0765	0.0747
10	0.0763	0.0805	0.0799	0.0805	0.0796
11	0.0661	0.0685	0.0678	0.0679	0.0676
12	0.0717	0.0681	0.0681	0.0678	0.0677
13	0.0631	0.0687	0.0692	0.0690	0.0693
14	0.0673	0.0685	0.0680	0.0685	0.0686
15	0.0793	0.0721	0.0707	0.0728	0.0713
16	0.0657	0.0696	0.0702	0.0697	0.0704
17	0.0859	0.0778	0.0759	0.0773	0.0759
18	0.0583	0.0626	0.0633	0.0618	0.0626
19	0.0619	0.0649	0.0647	0.0653	0.0647
20	0.0877	0.0923	0.0914	0.0917	0.0908

Area ID	Direct Est.	Comparison of CVs			
		FHM	CAR-FHM	YCM	CAR-YCM
1	0.168	0.107	0.099	0.107	0.100
2	0.127	0.105	0.104	0.097	0.093
3	0.135	0.116	0.106	0.110	0.097
4	0.102	0.084	0.076	0.079	0.072
5	0.158	0.094	0.076	0.105	0.083
6	0.113	0.086	0.080	0.086	0.081
7	0.124	0.099	0.096	0.106	0.101
8	0.102	0.085	0.076	0.081	0.073
9	0.158	0.119	0.105	0.117	0.105
10	0.121	0.087	0.086	0.086	0.084
11	0.141	0.118	0.108	0.109	0.105
12	0.196	0.119	0.109	0.130	0.116
13	0.168	0.115	0.108	0.111	0.108
14	0.206	0.126	0.125	0.136	0.133
15	0.121	0.101	0.087	0.094	0.083
16	0.127	0.101	0.097	0.103	0.097
17	0.124	0.107	0.100	0.105	0.096
18	0.155	0.143	0.136	0.134	0.130
19	0.154	0.135	0.134	0.128	0.128
20	0.103	0.086	0.085	0.083	0.082

References

- Arora, V., and Lahiri, P. (1997). On the superiority of the Bayesian method over the BLUP in small area estimation problems. *Statistica Sinica*, 7, 1053-1063.
- Bayarri, M.J., and Berger, J.O. (2000). P values for composite null models. *Journal of the American Statistical Association*, 95, 1127-1142.
- Béland, Y. (2002). Canadian Community Health Survey Methodological Overview. Health Report, Statistics Canada, Catalogue no. 82-003, 13, 3, 9-14.
- Besag, J., York, J. and Mollie, A. (1991). Bayesian image restoration, with two applications in spatial statistics (with discussion). *Annals of the Institute of Statistical Mathematics*, 43, 1-59.
- Best, N., Richardson, S. and Thomson, A. (2005). A comparison of Bayesian spatial models for disease mapping. *Statistical Methods in Medical Research*, 14, 35-39.
- Brown, G., Chambers, R., Heady, P. and Heasman, D. (2001). Evaluation of small area estimation methods – An application to unemployment estimates from the UK LFS. Proceedings: Symposium 2001, *Achieving Data Quality in a Statistical Agency: A Methodological Perspective*, CD-ROM, 1-10.
- Chip, S., and Greenberg, E. (1995). Understanding the Metropolitan-Hastings algorithm. *The American Statistician*, 49, 327-335.
- Cressie, N. (1990). Small area prediction of undercount using the general linear model. Proceedings: Symposium 1990, *Measurement and Improving Data Quality*, Statistics Canada, 93-105.
- Daniels, M.J., and Gatsonis, C. (1999). Hierarchical generalized linear models in the analysis of variations in health care utilization. *Journal of the American Statistical Association*, 94, 29-42.
- Datta, G.S., Lahiri, P., Maiti, T. and Lu, K.L. (1999). Hierarchical Bayes estimation of unemployment rates for the states of the U.S. *Journal of the American Statistical Association*, 94, 1074-1082.
- Dick, P. (1995). Modelling net undercoverage in the 1991 Canadian census. *Survey Methodology*, 21, 45-54.
- Fay, R.E., and Herriot, R.A. (1979). Estimation of income for small places: An application of James-Stein procedures to census data. *Journal of the American Statistical Association*, 74, 268-277.
- Gelfand, A.E. (1996). Model determination using sampling-based methods. In *Markov Monte Carlo in Practice* (Eds., W.R. Gilks, S. Richardson and D.J. Spiegelhalter), London: Chapman & Hall, 145-161.
- Gelman, A., Carlin, J.B., Stern, H.S. and Rubin, D.B. (2004). *Bayesian Data Analysis*, 2nd Edition. Chapman & Hall/CRC.
- Gelman, A., Meng, X.L. and Stern, H.S. (1996). Posterior predictive assessment of model fitness via realized discrepancies (with discussion). *Statistica Sinica*, 6, 733-807.
- Gelman, A., and Rubin, D.B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, 7, 457-472.
- Ghosh, M., Natarajan, K., Stroud, T.W.F. and Carlin, B.P. (1998). Generalized linear models for small area estimation. *Journal of the American Statistical Association*, 93, 273-282.
- Ghosh, M., Natarajan, K., Walter, L.A. and Kim, D.H. (1999). Hierarchical Bayes GLMs for the analysis of spatial data: An application to disease mapping. *Journal of Statistical Planning and Inference*, 75, 305-318.
- He, Z., and Sun, D. (2000). Hierarchical Bayes estimation of hunting success rates with spatial correlations. *Biometrics*, 56, 360-367.
- Jiang, J., and Lahiri, P. (2006). Mixed model prediction and small area estimation (with discussion). *Test*, 15, 1-96.
- Lele, S.R., Dennis, B. and Lutscher, F. (2007). Data cloning: Easy maximum likelihood estimation for complex ecological models using Bayesian Markov chain Monte Carlo methods. *Ecology Letters*, 10, 551-563.
- Lele, S.R., Nadeem, K. and Schmuland, B. (2010). Estimability and likelihood inference for generalized linear mixed models using data cloning. *Journal of the American Statistical Association*, 105, 1617-1625.
- Leroux, B.G., Lei, X. and Breslow, N. (1999). Estimation of disease rates in small areas: A new mixed model for spatial dependence. In *Statistical Models in Epidemiology, the Environment and Clinical Trials*, (Eds., M.E. Halloran and D. Berry). New York: Springer Verlag, 135-178.
- Liu, B., Lahiri, P. and Kalton, G. (2008). Hierarchical Bayes modeling of survey weighted small area proportions. Unpublished Manuscript.
- Maiti, T. (1998). Hierarchical Bayes estimation of mortality rates for disease mapping. *Journal of Statistical Planning and Inference*, 69, 339-348.
- Mohadjer, L., Rao, J.N.K., Liu, B., Krenzke, T. and van de Kerckhove, W. (2007). Hierarchical Bayes small area estimates of adult literacy using unmatched sampling and linking models. *Proceedings of the American Statistical Association, Section of Survey Method Research*.
- Moura, F.A.S., and Migon, H.S. (2002). Bayesian spatial models for small area proportions. *Statistical Modelling*, 2, 3, 183-201.
- MacNab, Y.C. (2003). Hierarchical Bayesian spatial modeling of small-area rates of non-rare disease. *Statistics in Medicine*, 22, 1761-1773.
- Maiti, T. (1998). Hierarchical Bayes estimation of mortality rates for disease mapping. *Journal of Statistical Planning and Inference*, 69, 339-348.
- Meng, X.L. (1994). Posterior predictive p value. *The Annals of Statistics*, 22, 1142-1160.
- Mollié, A. (1996). Bayesian mapping of disease. In *Markov Chain Monte Carlo in Practice*. London: Chapman and Hall, 359-379.
- Rao, J.N.K. (2003). *Small Area Estimation*. New York: John Wiley & Sons, Inc.
- Singh, A.C., Folsom, R.E., Jr. and Vaish, A.K. (2005). Small area modeling for survey data with smoothed error covariance structure via generalized design effects. Federal Committee on Statistical methods Conference proceedings, Washington, D.C., www.fcsm.gov.
- Singh, B.B., Shukla, G.K. and Kundu, D. (2005). Spatio-temporal models in small area estimation. *Survey Methodology*, 31, 183-195.

- Sinharay, S., and Stern, H.S. (2003). Posterior predictive model checking in hierarchical models. *Journal of Statistical Planning and Inference*, 111, 209-221.
- Spiegelhalter, D.J., Best, N., Carlin, B.P. and van de Linde, A. (2002). Bayesain measures of model complexity and fit. *Journal of Royal Statistical Society*, B, 64, 583-639.
- Souza, D.F., Moura, F.A.S. and Migon, H.S. (2009). Small area population prediction via hierarchical models. *Survey Methodology*, 35, 203-214.
- Wang, J., and Fuller, W. A. (2003). The mean square error of small area predictors constructed with estimated area variances. *Journal of the American Statistical Association*, 98, 716-723.
- You, Y. (2008a). An integrated modeling approach to unemployment rate estimation for sub-provincial areas of Canada. *Survey Methodology*, 34, 19-27.
- You, Y. (2008b). Small area estimation using area level models with model checking and applications. *Proceedings of Survey Methods Section, Statistical Society of Canada*.
- You, Y., and Chapman, B. (2006). Small area estimation using area level models and estimated sampling variances. *Survey Methodology*, 32, 97-103.
- You, Y., and Rao, J.N.K. (2000). Hierarchical Bayes estimation of small area means using multi-level models. *Survey Methodology*, 26, 173-181.
- You, Y., and Rao, J.N.K. (2002). Small area estimation using unmatched sampling and linking models. *The Canadian Journal of Statistics*, 20, 3-15.

Small area estimation under transformation to linearity

Hukum Chandra and Ray Chambers¹

Abstract

Small area estimation based on linear mixed models can be inefficient when the underlying relationships are non-linear. In this paper we introduce SAE techniques for variables that can be modelled linearly following a non-linear transformation. In particular, we extend the model-based direct estimator of Chandra and Chambers (2005, 2009) to data that are consistent with a linear mixed model in the logarithmic scale, using model calibration to define appropriate weights for use in this estimator. Our results show that the resulting transformation-based estimator is both efficient and robust with respect to the distribution of the random effects in the model. An application to business survey data demonstrates the satisfactory performance of the method.

Key Words: Sample survey; Survey estimation; Business surveys; Model calibration; Skewed data; Model-based direct estimation; Empirical best linear unbiased prediction.

1. Introduction

Commonly used methods for small area estimation (SAE) assume that a linear mixed model can be used to characterize the regression relationship between the survey variable Y and an auxiliary variable X in the small areas of interest. In particular, empirical best linear unbiased prediction (EBLUP), see Rao (2003, chapters 6 - 8) is typically based on a linear mixed model assumption. However, when the data are skewed, as is often the case in business surveys, the relationship between Y and X may not be linear in the original (raw) scale, but can be linear in a transformed scale, *e.g.*, the logarithmic (log) scale. In such cases we would expect estimation based on a linear mixed model for Y to be inefficient compared with one based on a similar model for a transformed version of Y . See Hidiroglou and Smith (2005). The use of transformations in inference has a long history, see for example Carroll and Ruppert (1988, chapter 4). Recently, Chen and Chen (1996) and Karlberg (2000a) have investigated the use of a 'transform to linearity' approach for regression estimation of survey variables that behave non-linearly. However, to the best of our knowledge there has been no application of this idea in SAE, even though economic theory (and casual observation) suggests that regression relationships in business survey data are typically multiplicative, and hence linear in the log scale.

In this paper we extend the model-based direct (MBD) estimation ideas described in Chandra and Chambers (2005, 2009) to the situation where the linear mixed model underpinning SAE holds on the log scale, using weights derived via model calibration (Wu and Sitter 2001). In doing so, we note that our approach easily generalises to

other monotone (*i.e.*, invertible) transformations. In contrast, extension of the EBLUP approach to where the data follow a linear mixed model under transformation is complicated. We also relax the usual normality assumption for the area effects in order to examine robustness with respect to this assumption.

In the following section we summarise the MBD approach to SAE under a linear mixed model. In section 3 we describe an alternative to the linear mixed model for skewed data which reduces to the linear mixed model under log transformation, and in section 4 we use a model-based perspective to motivate model calibrated estimation of population quantities where the underlying variable is linear after suitable transformation. In section 5 we bring these two ideas together, introducing the concept of a fitted value model derived from a linear mixed model in the transformed scale. We then use this fitted value model to specify survey weights for use in an MBD estimator in SAE. In section 6 we present empirical results from a number of simulation studies that contrast the proposed transformation-based MBD estimator with both the EBLUP and the 'usual' MBD estimator defined by fitting a linear mixed model to the data as well as with an indirect empirical predictor based on the same transformed scale linear mixed model. Section 7 concludes the paper with a discussion of outstanding issues.

Note that the approach taken in this article is model-based. Consequently all moments are evaluated with respect to a model for the population data. Also, all sample data are assumed to have been obtained via a non-informative sampling method, *e.g.*, probability sampling with inclusion probabilities defined by known model covariates.

1. Hukum Chandra, Indian Agricultural Statistics Research Institute, Library Avenue, PUSA Campus, New Delhi-110012, India. E-mail: hchandra@iasri.res.in; Ray Chambers, Centre for Statistical and Survey Methodology, University of Wollongong, Wollongong, NSW, 2522, Australia. E-mail: ray@uow.edu.au.

2. Model-based direct estimation for small areas

To start, we fix our notation. Let U denote a population of size N and let \mathbf{y}_U denote the N -vector of population values of a characteristic Y of interest. Suppose that our primary aim is estimation of the total $t_{Uy} = \sum_U y_j$ of these population values (or their mean $m_{Uy} = N^{-1} \sum_U y_j$). Let \mathbf{X} denote a p -vector of auxiliary variables that are related, in some sense, to Y and let \mathbf{x}_U denote the corresponding $N \times p$ matrix of population values these variables. We assume that the individual sample values of \mathbf{X} are known. The non-sample values of \mathbf{X} may not be individually known, but are assumed known at some aggregate level. At a minimum, we know the vector of population totals \mathbf{t}_{Ux} of the columns of \mathbf{X} .

Suppose that it is reasonable to assume that the regression of Y on \mathbf{X} in the population is linear, *i.e.*,

$$E(\mathbf{y}_U | \mathbf{x}_U) = \mathbf{x}_U \boldsymbol{\beta} \text{ and } \text{Var}(\mathbf{y}_U | \mathbf{x}_U) = \mathbf{v}_U \quad (1)$$

where \mathbf{v}_U is known up to a multiplicative constant. Given a sample s of size n from this population, we can partition

$$\mathbf{x}_U = \begin{bmatrix} \mathbf{x}_s \\ \mathbf{x}_r \end{bmatrix}$$

and

$$\mathbf{v}_U = \begin{bmatrix} \mathbf{v}_{ss} & \mathbf{v}_{sr} \\ \mathbf{v}_{rs} & \mathbf{v}_{rr} \end{bmatrix}$$

into their sample and non-sample components. Here $r = U - s$ denotes the population units that are not in sample. The vector of weights that defines the Best Linear Unbiased Predictor (BLUP) of t_{Uy} is then (Royall 1976; Valliant, Dorfman and Royall 2000, section 2.4)

$$\begin{aligned} \mathbf{w}_s^{\text{BLUP}} &= (\mathbf{w}_j^{\text{BLUP}}; j \in s) \\ &= \mathbf{1}_s + \mathbf{H}_s'(\mathbf{t}_{Ux} - \mathbf{t}_{sx}) + (\mathbf{I}_s - \mathbf{H}_s' \mathbf{x}_s') \mathbf{v}_{ss}^{-1} \mathbf{v}_{sr} \mathbf{1}_r \end{aligned} \quad (2)$$

where $\mathbf{H}_s = (\mathbf{x}_s' \mathbf{v}_{ss}^{-1} \mathbf{x}_s)^{-1} \mathbf{x}_s' \mathbf{v}_{ss}^{-1}$, \mathbf{I}_s is the identity matrix of order n , \mathbf{t}_{sx} is the vector of sample totals of \mathbf{X} and $\mathbf{1}_s$ ($\mathbf{1}_r$) denotes a vector of ones of size n ($N - n$).

We now assume that the target population U of size N can be partitioned into D non-overlapping small areas or domains, each of size N_i , $i = 1, \dots, D$, such that $N = \sum_{i=1}^D N_i$. Given a sample s of size n units is drawn from this population, we shall assume that a sub-sample s_i of size n_i units is drawn from area i , with $n = \sum_{i=1}^D n_i$. Note that we assume that all small areas are sampled and that there is at least one sample unit in each small area of interest.

As noted in section 1, linear mixed models are often used in SAE. Such models can be written in the form

$$\mathbf{y}_U = \mathbf{x}_U \boldsymbol{\beta} + \mathbf{g}_U \mathbf{u} + \mathbf{e}_U \quad (3)$$

where \mathbf{u} is a random vector of so-called area effects, \mathbf{e}_U is a population N -vector of random individual effects and \mathbf{g}_U is a known matrix. In general, area effects are vector-valued, so $\mathbf{u}' = (\mathbf{u}_1' \mathbf{u}_2' \dots \mathbf{u}_D')$ and $\mathbf{g}_U = \text{diag}\{\mathbf{g}_i; i = 1, \dots, D\}$, where \mathbf{g}_i is of dimension $N_i \times q$. The area specific effects $\{\mathbf{u}_i; i = 1, \dots, D\}$ are assumed to be independent and identically distributed realisations of a random vector of dimension q with zero mean and covariance matrix Σ_u . Similarly, the scalar individual effects making up \mathbf{e}_U are assumed to be independent and identically distributed realisations of a random variable with zero mean and variance σ_e^2 , with area and individual effects mutually independent. The parameters $\theta = (\Sigma_u, \sigma_e^2)$ are typically referred to as the variance components of (3).

Given the values of the variance components, it is straightforward to see that (3) is just a special case of the general linear model (1) that underpins the BLUP weights (2). In particular, under (3)

$$\begin{aligned} \mathbf{v}_{ss} &= \text{diag}\{\mathbf{v}_{iss}; i = 1, \dots, D\} \\ &= \text{diag}\{\mathbf{g}_i \Sigma_u \mathbf{g}_i' + \sigma_e^2 \mathbf{I}_{N_i}; i = 1, \dots, D\} \end{aligned} \quad (4)$$

and

$$\begin{aligned} \mathbf{v}_{sr} &= \text{diag}\{\mathbf{v}_{isr}; i = 1, \dots, D\} \\ &= \text{diag}\{\mathbf{g}_i \Sigma_u \mathbf{g}_{ir}'; i = 1, \dots, D\}. \end{aligned} \quad (5)$$

Here \mathbf{g}_{is} and \mathbf{g}_{ir} denote the restriction of \mathbf{g}_i to sampled and non-sampled units in area i respectively. Given estimated values $\hat{\theta} = (\hat{\Sigma}_u, \hat{\sigma}_e^2)$ of the variance components we can substitute these in (4) and (5) to obtain estimates $\hat{\mathbf{v}}_{ss}$ and $\hat{\mathbf{v}}_{sr}$ of \mathbf{v}_{ss} and \mathbf{v}_{sr} respectively, and therefore compute 'empirical' BLUP weights, or EBLUP weights for the population total of Y as

$$\begin{aligned} \mathbf{w}_s^{\text{EBLUP}} &= (\mathbf{w}_{ij}^{\text{EBLUP}}; j \in s_i; i = 1, \dots, D) \\ &= \mathbf{1}_s + \hat{\mathbf{H}}_s'(\mathbf{t}_{Ux} - \mathbf{t}_{sx}) \\ &\quad + (\mathbf{I}_s - \hat{\mathbf{H}}_s' \mathbf{x}_s') \hat{\mathbf{v}}_{ss}^{-1} \hat{\mathbf{v}}_{sr} \mathbf{1}_r \end{aligned} \quad (6)$$

where $\hat{\mathbf{H}}_s = (\mathbf{x}_s' \hat{\mathbf{v}}_{ss}^{-1} \mathbf{x}_s)^{-1} \mathbf{x}_s' \hat{\mathbf{v}}_{ss}^{-1}$. Note that we now use a double index of ij to differentiate between population units in different areas.

The MBD estimator for the mean m_{iy} of Y in area i (Chandra and Chambers 2005, 2009) based on the EBLUP weights for the total (6) is simply the corresponding weighted average of the sample values of Y in area i ,

$$\hat{m}_{iy}^{\text{HJ-LinMBD}} = \left\{ \sum_{j \in s_i} w_{ij}^{\text{EBLUP}} \right\}^{-1} \sum_{j \in s_i} w_{ij}^{\text{EBLUP}} y_{ij}. \quad (7)$$

Note that (7) is *not* the EBLUP for m_{iy} under (3). This is (see Rao 2003, section 6.2.3)

$$\begin{aligned} \hat{m}_{iy}^{\text{HT-LinEBLUP}} &= \hat{E}\{m_{iy} | \mathbf{y}_{is}, \mathbf{x}_{is}, \mathbf{x}_{ir}\} \\ &= N_i^{-1} \left[\sum_{j \in s_i} y_j + \mathbf{1}'_{ir} \left\{ \mathbf{x}_{ir} \hat{\beta} + \hat{\mathbf{v}}_{irs} \hat{\mathbf{v}}_{iss}^{-1} (\mathbf{y}_{is} - \mathbf{x}_{is} \hat{\beta}) \right\} \right] \\ &= N_i^{-1} \left[n_i \bar{y}_{is} + (N_i - n_i) \right. \\ &\quad \left. \left\{ \bar{\mathbf{x}}'_{ir} \hat{\beta} + \bar{\mathbf{g}}'_{ir} \hat{\Sigma}_u \mathbf{g}'_{is} (\mathbf{g}_{is} \hat{\Sigma}_u \mathbf{g}'_{is} + \hat{\sigma}_e^2 \mathbf{I}_{is})^{-1} (\mathbf{y}_{is} - \mathbf{x}_{is} \hat{\beta}) \right\} \right]. \quad (8) \end{aligned}$$

Here \hat{E} denotes the expectation operator under (3) with unknown parameters replaced by estimates, \mathbf{x}_{is} and \mathbf{x}_{ir} are the matrices of sample and non-sample values of \mathbf{X} in area i , \mathbf{y}_{is} is the vector of sample values of Y in the same area, $\hat{\beta}$ is the 'empirical' BLUE of β , $\hat{\mathbf{v}}_{irs}$ is the transpose of the estimated value of \mathbf{v}_{irs} with $\hat{\mathbf{v}}_{iss}$ the corresponding estimate of \mathbf{v}_{iss} , see (4) and (5), and $\mathbf{1}_{ir}$ is a vector of ones of length $N_i - n_i$. Note that the last expression on the right hand side of (8) follows directly by substitution of (4) and (5), with $\bar{\mathbf{x}}_{ir}$ and $\bar{\mathbf{g}}_{ir}$ denoting the column vectors of order p and q defined by averaging the columns of \mathbf{x}_{ir} and \mathbf{g}_{ir} respectively. Like the EBLUP (8), the estimator (7) is a weighted function of all the sample values. Note that under random intercept specification of (3), (8) reduces to the expression (7.2.39) in Rao (2003, section 7.2).

Mean squared error (MSE) estimation for (8) is usually carried out using the theory described in Prasad and Rao (1990). Although this MSE estimator is somewhat complicated, it works well under (3). However, when (3) fails it can be misleading. It is also inadequate as an estimator of the repeated sampling MSE of (8), as has been pointed out by Longford (2007). In contrast, MSE estimation for (7) is quite straightforward. This is because if one treats the weights defining this estimator as fixed, then it is a linear estimator of a domain mean, and so its prediction variance V_i under (1) can be estimated using well-known methods (see Royall and Cumberland 1978). Since in general the EBLUP weights for the total (6) are not 'locally calibrated' (i.e., they do not reproduce the area i mean $\bar{\mathbf{x}}_i$ of \mathbf{X}), (7) has a bias B_i under (1). A simple plug-in estimate of this bias is the difference between (7) and $\bar{\mathbf{x}}'_i \hat{\beta}$. The final MSE estimator used with (7) is therefore defined by summing the estimate of V_i and the square of this estimate of B_i . This method of MSE estimation has been empirically demonstrated to have good model-based as well as repeated sampling properties. See Chandra and Chambers (2005, 2009), Chambers and Tzavidis (2006), Chandra, Salvati and

Chambers (2007) and Tzavidis, Salvati, Pratesi and Chambers (2008).

3. Small area estimation under transformation

In this section we extend the MBD approach to SAE when the underlying regression relationships are non-linear. In doing so, we shall focus on the important case where the population values of Y follow a non-linear model in their original (raw) scale, but their logarithms can be modelled linearly. The extension to other 'transform to linear' models is straightforward.

Without loss of generality, suppose that both Y and X are scalar and strictly positive, with skewed population marginal distributions and clear evidence of non-linearity in their relationship, e.g., as in many business surveys applications. Furthermore, a linear mixed model is appropriate for characterising how the regression of $\log(Y)$ on $\log(X)$ varies between the small areas. That is, for $i = 1, \dots, D$; $j = 1, \dots, N_i$ we have

$$l_{ij} = \log(y_{ij}) = \beta_0 + \beta_1 \log(x_{ij}) + \mathbf{g}'_{ij} \mathbf{u}_i + e_{ij} \quad (9)$$

where y_{ij} and x_{ij} are the values of Y and X respectively for population unit j in small area i , \mathbf{g}_{ij} denotes a 'contextual' covariate of dimension q , \mathbf{u}_i denotes a random effect for area i also of dimension q and e_{ij} is a scalar individual random effect. As usual with this type of model, we assume that all random effects are normally distributed and mutually uncorrelated, with zero expected values, $\text{Var}(\mathbf{u}_i) = \Sigma_u$ and $\text{Var}(e_{ij}) = \sigma_e^2$. Here Σ_u is the $q \times q$ matrix of covariances for the random effects. Note that $\text{Var}(l_{ij} | x_{ij}) = v_{ij} = \mathbf{g}'_{ij} \Sigma_u \mathbf{g}_{ij} + \sigma_e^2$ and $\text{Cov}(l_{ij}, l_{ik} | x_{ij}, x_{ik}, \mathbf{g}_{ij}, \mathbf{g}_{ik}) = v_{ijk} = \mathbf{g}'_{ij} \Sigma_u \mathbf{g}_{ik}$ under (9).

Given sample values of y_{ij} , x_{ij} and \mathbf{g}_{ij} , standard methods of estimation (e.g., ML or REML, see Harville 1977) can be used to estimate the parameters of (9). Let $\hat{\Sigma}_u$ and $\hat{\sigma}_e^2$ denote the resulting estimates of the variance components of this linear mixed model. The estimate of $\beta = (\beta_0 \ \beta_1)'$ is then

$$\hat{\beta} = \left(\sum_i \mathbf{d}'_i \hat{\mathbf{v}}_{iss}^{-1} \mathbf{d}_i \right)^{-1} \left(\sum_i \mathbf{d}'_i \hat{\mathbf{v}}_{iss}^{-1} \mathbf{l}_i \right) \quad (10)$$

where $\hat{\mathbf{v}}_{iss}$, \mathbf{d}_{is} and \mathbf{l}_{is} are the sample components of $\hat{\mathbf{v}}_i = [\hat{v}_{ijk}] = \mathbf{g}'_i \hat{\Sigma}_u \mathbf{g}'_i + \hat{\sigma}_e^2 \mathbf{I}_i$, $\mathbf{d}_i = [\mathbf{d}_{ijk}] = [\mathbf{1}_i \log(\mathbf{x}_i)]$ and $\mathbf{l}_i = (l_{ij}; j = 1, \dots, N_i)$ respectively. Here \mathbf{g}_i is the $N_i \times q$ matrix defined by the covariates \mathbf{g}_{ij} in area i , \mathbf{I}_i is the identity matrix of order N_i , $\mathbf{1}_i$ denotes a vector of ones of dimension N_i and $\log(\mathbf{x}_i)$ denotes the vector of N_i values of $\log(X)$ in area i .

Note that when the variance components Σ_u and σ_e^2 are known, (10) is the BLUE for β . Consequently, $E(\hat{\beta}) \approx \beta$

and $\text{Var}(\hat{\beta}) \approx (\sum_i \mathbf{d}'_{is} \hat{\mathbf{V}}_{iss}^{-1} \mathbf{d}_{is})^{-1}$. Put $\hat{\phi}_i = (\hat{\phi}_{ij}) = \mathbf{d}_i \hat{\beta}$. Then $E(\hat{\phi}_i) \approx \mathbf{d}_i \beta$ and $\text{Var}(\hat{\phi}_i) = \mathbf{A}_i = [a_{ijk}] \approx \mathbf{d}_i (\sum_g \mathbf{d}'_{gs} \hat{\mathbf{V}}_{gss}^{-1} \mathbf{d}_{gs})^{-1} \mathbf{d}'_i$, where $a_{ijk} = \mathbf{d}'_{ij} \text{Var}(\hat{\beta}) \mathbf{d}_{ik} \rightarrow 0$ as $n \rightarrow \infty$.

Our aim is to use the log scale linear mixed model (9) for estimation of the small area means m_{iv} . In particular, we use model calibration (Wu and Sitter 2001) based on this model to develop sample weights for use in the MBD estimator (7) of this quantity.

4. Model calibrated weighting

Model calibration was introduced by Wu and Sitter (2001) as a model-assisted method of calibrated weighting when the underlying regression relationship is non-linear. Here we provide a model-based perspective on the method, as a precursor to using it for constructing weights for use in an MBD estimator in a similar situation.

Suppose that the underlying population model is non-linear, with the relationship between Y and \mathbf{X} in the population of form

$$E(y_j | \mathbf{x}_j) = h(\mathbf{x}_j; \eta) \text{ and } \text{Var}(y_j | \mathbf{x}_j) = \sigma_j^2. \quad (11)$$

Here $j = 1, \dots, N$, η (typically vector-valued) and σ_j^2 are unknown model parameters and the mean function $h(\mathbf{x}_j; \eta)$ is a known function of \mathbf{x}_j and η . We also assume that population units are mutually uncorrelated given their respective values of \mathbf{X} . Note that (11) is quite general, and includes linear, non-linear, and generalized linear models as special cases. In this situation, Wu and Sitter (2001) define the model-calibrated estimator of the population total t_{Uy} as $\hat{t}_{Uy}^{mc} = \sum_{j \in s} w_j^{mc} y_j$, where the vector of weights $\mathbf{w}_s^{mc} = (w_j^{mc})$ is chosen to minimise an appropriately chosen measure of the distance from \mathbf{w}_s^{mc} to the vector of Horvitz-Thompson weights $\mathbf{w}_s^\pi = (\pi_j^{-1})$, subject to the model calibration constraints

$$\sum_{j \in s} w_j^{mc} = N$$

and

$$\sum_{j \in s} w_j^{mc} h(\mathbf{x}_j; \hat{\eta}_\pi) = \sum_{j \in U} h(\mathbf{x}_j; \hat{\eta}_\pi) \quad (12)$$

with $\hat{\eta}_\pi$ a design consistent estimator of η . Note that unlike standard calibration, the constraints (12) require that we know the individual population values of \mathbf{X} . The key idea behind this approach is that provided (11) fits reasonably, then y_j is (at least approximately) a linear function of its fitted value $h(\mathbf{x}_j; \hat{\eta}_\pi)$ under this model and so we can carry out linear estimation using these fitted values as auxiliary information.

A model-based perspective on model calibration can be developed as follows. Let $\hat{\eta}$ denote a 'model-efficient' estimator of η in (11), e.g., its maximum likelihood (ML) estimator, with associated fitted values $h(\mathbf{x}_j; \hat{\eta})$. In general, these fitted values will not be unbiased. They will also be correlated. However, there will still be a systematic relationship between the actual values of Y and their corresponding fitted values that we can approximate. Although there is nothing to stop us looking at more complex approximations, a linear model for the relationship between the population values y_j and the fitted values $\hat{y}_j = h(\mathbf{x}_j; \hat{\eta})$ seems a reasonable starting point. We therefore replace the non-linear model (11) by the linear model

$$E(y_j | \hat{y}_j) = \alpha_0 + \alpha_1 \hat{y}_j$$

and

$$(13)$$

$$\text{Cov}(y_j, y_k | \hat{y}_j, \hat{y}_k) = \omega_{jk}.$$

We refer to (13) as the 'fitted value' model corresponding to (11). Let \mathbf{J}_U denote the population 'design matrix' under (13), i.e., $\mathbf{J}_U = [\mathbf{1}_U \hat{\mathbf{y}}_U]$, where $\mathbf{1}_U$ denotes the unit vector of size N and $\hat{\mathbf{y}}_U = (\hat{y}_j; j = 1, \dots, N)$, and put $\Omega_U = [\omega_{jk}; j = 1, \dots, N; k = 1, \dots, N]$. We can then partition \mathbf{J}_U and Ω_U according to sample (s) and non-sample (r) units as

$$\mathbf{J}_U = \begin{bmatrix} \mathbf{J}_s \\ \mathbf{J}_r \end{bmatrix}$$

and

$$\Omega_U = \begin{bmatrix} \Omega_{ss} & \Omega_{sr} \\ \Omega_{rs} & \Omega_{rr} \end{bmatrix},$$

and hence write down the weights that define the BLUP of t_{Uy} under (13). These are the model-based model-calibrated weights

$$\mathbf{w}^{mbmc} = (w_j^{mbmc}; j \in s)$$

$$= \mathbf{1}_s + \mathbf{H}'_{cm} (\mathbf{J}'_U \mathbf{1}_U - \mathbf{J}'_s \mathbf{1}_s) + (\mathbf{I}_s - \mathbf{H}'_{cm} \mathbf{J}'_s) \Omega_{ss}^{-1} \Omega_{sr} \mathbf{1}_r \quad (14)$$

where $\mathbf{H}_{mc} = (\mathbf{J}'_s \Omega_{ss}^{-1} \mathbf{J}_s)^{-1} \mathbf{J}'_s \Omega_{ss}^{-1}$. Clearly, these weights are model-calibrated since $\sum_{j \in s} w_j^{mbmc} = N$ and $\sum_{j \in s} w_j^{mbmc} \hat{y}_j = \sum_{j \in U} \hat{y}_j$. However, unlike the linear model EBLUP weights (2), they are *not* calibrated on \mathbf{X} . In practice, the components of Ω_U will not be known and will need to be estimated. When these estimates are substituted in (14), we obtain the empirical version \mathbf{w}^{embmc} of these model-calibrated weights.

5. Model calibrated weighting for small area estimation

We now use model calibration based on the log scale linear mixed model (9) to obtain sample weights for use in the MBD estimator (7). From the development in the previous section it can be seen that this requires us to first specify a fitted value model (13) for Y based on (9), *i.e.*, we need to calculate appropriate fitted values \hat{y}_{ij} as well as estimates $\hat{\omega}_{ijk}$ of $\omega_{ijk} = \text{Cov}(y_{ij}, y_{ik} | x_{ij}, x_{ik}, \mathbf{g}_{ij}, \mathbf{g}_{ik})$ under (9). The sample weights to use in the MBD estimator (7) are then given by (14).

A simple method of defining fitted values \hat{y}_{ij} under (9) is one where parameter estimates derived under this model are used to obtain predicted values on the log scale which are then back-transformed. Unfortunately, as is well known, this approach is biased. We therefore develop the first and second order moments of an appropriate bias-corrected fitted value model based on (9). Let \mathbf{x}_s and \mathbf{g}_s denote the sample values of x_{ij} and \mathbf{g}_{ij} respectively. Under (9),

$$E(y_{ij} | x_{ij}, \mathbf{g}_{ij}) = E\{e^{l_{ij}} | x_{ij}, \mathbf{g}_{ij}\} = e^{\phi_{ij} + v_{ij}/2} \\ \neq E(e^{\hat{\phi}_{ij} + \hat{v}_{ij}/2} | \mathbf{x}_s, \mathbf{g}_s) = E(\hat{y}_{ij} | x_{ij}, \mathbf{g}_{ij})$$

so the usual bias correction that makes use of the fact that the conditional distribution of y_{ij} is lognormal is inadequate. Let $\hat{\eta}_{ij} = (\hat{\beta}, \hat{v}_{ij})'$ be an estimate of $\eta_{ij} = (\beta, v_{ij})'$ such that $E(\hat{\eta}_{ij} - \eta_{ij}) \approx 0$ for large n . Put $z(\eta_{ij}) = e^{\phi_{ij} + v_{ij}/2}$. Using a second order Taylor series approximation we can write

$$z(\hat{\eta}_{ij}) \approx z(\eta_{ij}) + (\hat{\eta}_{ij} - \eta_{ij})' z^{(1)}(\eta_{ij}) \\ + \frac{1}{2} (\hat{\eta}_{ij} - \eta_{ij})' z^{(2)}(\eta_{ij}) (\hat{\eta}_{ij} - \eta_{ij})$$

and so

$$E\{z(\hat{\eta}_{ij})\} \approx z(\eta_{ij}) \\ + \frac{1}{2} \text{tr}[E\{z^{(2)}(\eta_{ij})(\hat{\eta}_{ij} - \eta_{ij})(\hat{\eta}_{ij} - \eta_{ij})'\}].$$

Here

$$z^{(1)}(\eta_{ij}) = \left(\mathbf{d}_{ij}' e^{\phi_{ij} + v_{ij}/2} \quad \frac{1}{2} e^{\phi_{ij} + v_{ij}/2} \right)'$$

and

$$z^{(2)}(\eta_{ij}) = \begin{pmatrix} \mathbf{d}_{ij}' \mathbf{d}_{ij}' e^{\phi_{ij} + v_{ij}/2} & \frac{1}{2} \mathbf{d}_{ij}' e^{\phi_{ij} + v_{ij}/2} \\ \frac{1}{2} \mathbf{d}_{ij}' e^{\phi_{ij} + v_{ij}/2} & \frac{1}{4} e^{\phi_{ij} + v_{ij}/2} \end{pmatrix}$$

are the vector and matrix respectively containing the first and second order derivatives of $z(\eta_{ij})$ with respect to η_{ij} . Since the asymptotic covariance between ML (or REML) estimators of the fixed and variance components of a linear mixed model is zero (McCulloch and Searle 2001, chapter 2, pages 40 - 45), the covariance between $\hat{\beta}$ and \hat{v}_{ij} will be negligible. It follows that

$$\text{tr}[E\{z^{(2)}(\eta_{ij})(\hat{\eta}_{ij} - \eta_{ij})(\hat{\eta}_{ij} - \eta_{ij})'\}] \\ = \text{tr}[z^{(2)}(\eta_{ij})E\{(\hat{\eta}_{ij} - \eta_{ij})(\hat{\eta}_{ij} - \eta_{ij})'\}] \\ \approx e^{\phi_{ij} + \frac{v_{ij}}{2}} \left[\mathbf{d}_{ij}' \left(\sum_g \mathbf{d}_{gs}' \hat{\mathbf{v}}_{gss}^{-1} \mathbf{d}_{gs} \right)^{-1} \mathbf{d}_{ij} + \frac{1}{4} \text{Var}(\hat{v}_{ij}) \right] \\ = E(y_{ij} | x_{ij}, \mathbf{g}_{ij}) \left[\hat{a}_{ij} + \frac{1}{4} \text{Var}(\hat{v}_{ij}) \right]$$

where $\hat{a}_{ij} = \mathbf{d}_{ij}' \hat{\mathbf{V}}(\hat{\beta}) \mathbf{d}_{ij}$ and $\hat{\mathbf{V}}(\hat{\beta}) = (\sum_i \mathbf{d}_{is}' \hat{\mathbf{v}}_{iss}^{-1} \mathbf{d}_{is})^{-1}$ is the usual estimator of $\text{Var}(\hat{\beta})$. Our fitted values are therefore defined by the second order bias corrected estimator of $E(y_{ij} | x_{ij}, \mathbf{g}_{ij})$,

$$\hat{y}_{ij} = h(\mathbf{d}_{ij}; \hat{\eta}_{ij}) = \hat{k}_{ij}^{-1} e^{\hat{\phi}_{ij} + \hat{v}_{ij}/2} \quad (15)$$

where

$$\hat{k}_{ij} = 1 + \frac{1}{2} \left\{ \hat{a}_{ij} + \frac{1}{4} \hat{\mathbf{V}}(\hat{v}_{ij}) \right\}$$

and $\hat{\mathbf{V}}(\hat{v}_{ij})$ is the estimated asymptotic variance of \hat{v}_{ij} . Under ML and REML estimation of the variance components of (9), this estimated asymptotic variance is obtained from the inverse of the relevant information matrix. Note that the bias adjustment of Karlberg (2000a) is a special case of (15).

In order to use (14) to define model-based model-calibrated sample weights, we also need estimates of the second order moments of the population values of Y given these fitted values. The conditional moments ω_{ijk} are a first order approximation to these moments. In particular, given normal random effects

$$\omega_{ijk} = e^{(\phi_{ij} + \phi_{ik}) + (v_{ij} + v_{ik})/2} (e^{v_{ijk}} - 1) \quad (16)$$

Our estimate $\hat{\omega}_{ijk}$ of ω_{ijk} is obtained by substituting $\hat{\phi}_{ij}$ and \hat{v}_{ijk} for ϕ_{ij} and v_{ijk} in (16).

The empirical model-based model-calibrated weights (14) corresponding to the fitted value model defined by (15) and (16) are

$$\mathbf{w}^{embmc} = (\mathbf{w}_{ij}^{embmc}; j \in S_i; i = 1, \dots, D) \\ = \mathbf{1}_s + \hat{\mathbf{H}}_{mc}' (\mathbf{J}_U' \mathbf{1}_U - \mathbf{J}_s' \mathbf{1}_s) \\ + (\mathbf{I}_s - \hat{\mathbf{H}}_{mc}' \mathbf{J}_s') \hat{\Omega}_{ss}^{-1} \hat{\Omega}_{sr} \mathbf{1}_r. \quad (17)$$

Here $\mathbf{J}_U = [\mathbf{1}_U \ \hat{\mathbf{y}}_U]$, so

$$\mathbf{J}'_U \mathbf{1}_U - \mathbf{J}'_s \mathbf{1}_s = \begin{pmatrix} N - n \\ \sum_i \sum_{j \in r_i} \hat{y}_{ij} \end{pmatrix},$$

and $\hat{\mathbf{H}}_{mc} = (\mathbf{J}'_s \hat{\Omega}_{ss}^{-1} \mathbf{J}_s)^{-1} \mathbf{J}'_s \hat{\Omega}_{ss}^{-1}$. Also $\hat{\Omega}_{ss} = \text{diag}\{\hat{\Omega}_{iss}; i = 1, \dots, D\}$ and $\hat{\Omega}_{sr} = \text{diag}\{\hat{\Omega}_{isr}; i = 1, \dots, D\}$, where $\hat{\Omega}_{iss}$ and $\hat{\Omega}_{isr}$ are defined by the sample/non-sample decomposition of $\hat{\Omega}_i$. For example, when (9) corresponds to a random intercepts specification, $\hat{y}_{ijk} = \hat{\sigma}_u^2 + \hat{\sigma}_e^2 I(j = k)$ and so the components of $\hat{\Omega}_i$ are

$$\hat{\omega}_{ijk} = e^{\hat{\phi}_{ij} + \hat{\phi}_{ik} + \hat{\sigma}_u^2 + \hat{\sigma}_e^2} [e^{\hat{\sigma}_e^2} \{1 + I(j = k)(e^{\hat{\sigma}_e^2} - 1)\} - 1].$$

The development so far has assumed normality of log-scale random effects. However, there is no good reason (beyond convenience) to assume that with skewed data these random area effects should be normal. One alternative, given a scalar area effect in (9), is to assume that the random effects in this model are drawn from the *gamma* family of distributions. From the properties of this distribution and using binomial and exponential expansions (ignoring higher order terms) we can show that $E(y_{ij} | x_{ij}, \mathbf{g}_{ij}) \approx e^{\phi_{ij} + \nu_{ij}/2} = z(\eta_{ij})$ as in the normal case. This indicates that an MBD estimator based on the model-based model-calibrated weights (17) should be robust with respect to the distribution of the random effects in (9).

Finally, we consider definition of the MBD estimator itself. As noted in section 2, this estimator is just the weighted average of the sample Y -values in an area. However, use of such a weighted average pre-supposes that the weights are reasonably close to being ‘locally calibrated on N_i ’, i.e., when summed over the sample units in small area i we obtain a value that is not too different from the actual small area population size N_i . This property usually holds if the weights are the EBLUP weights for the total (6) defined by a linear mixed model for Y . It does not necessarily hold for the model-based model-calibrated weights (17). Consequently, we consider two specifications for the MBD estimator given these weights. The first, which we refer to as a ‘Hájek specification’, is just the weighted average (7), with weights defined by (17). The second, which we refer to as a ‘Horvitz-Thompson specification’, replaces the denominator in (7) by the actual value of N_i . That is, the two types of MBD estimator under model-based model-calibrated weighting that we consider are

$$\hat{m}_{iy}^{\text{HJ-TrMBD}} = \left\{ \sum_{j \in s_i} w_{ij}^{\text{embmc}} \right\}^{-1} \sum_{j \in s_i} w_{ij}^{\text{embmc}} y_{ij} \quad (18)$$

and

$$\hat{m}_{iy}^{\text{HT-TrMBD}} = N_i^{-1} \sum_{j \in s_i} w_{ij}^{\text{embmc}} y_{ij}. \quad (19)$$

Alternatively we can adopt a prediction-based approach to obtain an alternative indirect predictor for the small area mean under the log-transformed model (9). Our approach extends that of Karlberg (2000a). In this case, assuming model (9) holds, we predict each nonsample Y in small area i and then sum these predictions. Note that we need to correct for bias following back-transformation to the raw scale when calculating these predicted values for the nonsample Y . Under model (9), the resulting empirical predictor for the mean m_{iy} of Y in area i (denoted TrEP) can be defined as

$$\hat{m}_{iy}^{\text{TrEP}} = N_i^{-1} \left\{ \sum_{j \in s_i} y_{ij} + \sum_{j \in r_i} \hat{y}_{ij} \right\}, \quad (20)$$

where \hat{y}_{ij} is given by (15).

Estimation of the MSE of (18) and (19) is carried out in the usual way for MBD estimators, i.e., via the MSE estimation approach described in section 2. Estimation of the MSE of (20) is not straightforward since this predictor is a non-linear function of Y values. We do not pursue this issue in this paper.

6. An empirical evaluation

In this section we provide empirical results on the comparative performances of five different methods of SAE. These are the two ‘transformation-based’ MBD estimators (18) and (19), both based on the model-based model-calibrated weights (17) and denoted by HJ-TrMBD and HT-TrMBD respectively; the log-transformation based predictor (20) under model (9), denoted TrEP, the ‘standard’ MBD estimator (7) based on the linear mixed model (3) and the empirical EBLUP weights for the total (6), which we denote by HJ-LinMBD to emphasise that it is a Hájek-type weighted mean based on weights derived under a linear mixed model; and the EBLUP (8) derived under the same linear mixed model, which we denote HT-LinEBLUP. Note that the MSEs for all three MBD estimators were estimated using the method described in section 2, while the MSE of HT-LinEBLUP was estimated using the method described in Prasad and Rao (1990). Note that we have not considered estimation of the MSE of TrEP.

Our empirical results are based on two types of simulation studies. The first type used model-based simulation to generate artificial population and sample data. That is, at each simulation population data were first generated under the model and a single sample was then taken from this simulated population by stratified simple random sampling without replacement with small area as strata. These data were then used to compare the performances of the different estimators. In section 6.1 we present the results from these model-based simulations. We carried out two

sets of model-based simulations. In the first set of simulations (Set A), we investigated the performance of these estimators given population data generated using the log-scale linear mixed model (9). In second set of simulations (Set B), we examined the robustness of these estimators to misspecification of this model. The second type of simulation study was design-based. In section 6.2 we describe design-based simulations. Here we evaluated these estimators in the context of repeated sampling from a real population using realistic sampling methods. That is, real survey data were first used to simulate a population, and this fixed population was then repeatedly sampled according to a pre-specified design. In particular, the sample design used was stratified random sampling with strata corresponding to the small areas of interest and with stratum allocations set to the small area sample sizes in the original datasets.

Four measures of estimator performance were computed using the various estimates generated in these simulation studies. They were the relative bias (RB) and the relative root mean squared error (RRMSE) of these estimates, together with the coverage rate and average width of the nominal 95 per cent confidence intervals based on them. In Tables 2 to 4 these measures are presented as averages over the small areas of interest.

6.1 The model-based simulation study

Model-based simulations are a common way of illustrating the sensitivity of an estimation procedure to variation in assumptions about the structure of the population of interest. Here we fixed the population size at $N = 15,000$ and randomly generated the small area population sizes $N_i, i = 1, \dots, D = 30$ so that $\sum_i N_i = N$. We used an overall sample size of $n = 600$ with small area sample sizes set so that they were proportional to the corresponding small area population sizes. These area-specific population and sample sizes were kept fixed in all our simulations. The population and sample sizes are given in Table 1a.

Table 1a
Area specific population (N_i) and sample (n_i) sizes for model-based simulation

Area	1	2	3	4	5	6	7	8	9	10
N_i	525	538	510	468	526	484	516	458	529	518
n_i	21	22	20	19	21	19	21	19	21	21
Area	11	12	13	14	15	16	17	18	19	20
N_i	502	524	509	484	487	459	542	498	512	500
n_i	20	21	20	19	19	18	22	20	20	20
Area	21	22	23	24	25	26	27	28	29	30
N_i	497	492	443	506	513	536	506	495	463	460
n_i	20	20	18	20	21	21	20	20	19	18

In Set A of our model-based simulations the population values y_{ij} were generated using the multiplicative model $y_{ij} = 5.0x_{ij}^\beta u_i e_{ij} (j = 1, \dots, N_i; i = 1, \dots, 30)$, with random samples then taken from each small area. Here the values of x_{ij} were independently drawn from the log-normal distribution $\log(x_{ij}) \sim N(6, \sigma_x^2)$, with the individual effects and area effects independently drawn as $\log(e_{ij}) \sim N(0, \sigma_e^2)$ and $\log(u_i) \sim N(0, \sigma_u^2)$ respectively. The population values of x were re-generated in each simulation. In particular, in each simulation we first generated the values of x 's for a population of size N and then randomly assigned these values to different areas of sizes N_i . The values of σ_e and σ_u were chosen so that the intra-area correlation in the population varied between 0.20 and 0.25. Table 1b shows the six different sets of parameter values that were used in Set A. These ensured that the simulated populations contained a wide range of variation. For each generated population and for each area i we selected a simple random sample (without replacement) of size n_i , leading to an overall sample size of $n = 600$. The sample values of y and the population values of x obtained in each simulation were then used to estimate the small area means. That is, using the sample data in each case, parameter values were estimated using the *lme* function in R (Bates and Pinheiro 1998), and estimates for the small area means then calculated, along with appropriate nominal 95% confidence intervals. The process of generating population and sample data, estimation of parameters and calculation of small area estimates was independently replicated 1,000 times. The results from this part of the simulation study are shown in Table 2.

Table 1b
Population specifications for model-based simulation Set A

Parameter Set	β	σ_u	σ_e	σ_x
1	0.5	0.30	0.50	3.00
2	0.8	0.35	0.60	2.50
3	1.0	0.40	0.70	2.25
4	1.3	0.45	0.80	1.75
5	1.5	0.50	0.90	1.50
6	2.0	0.60	1.00	1.20

In Set B of the model-based simulations, population data were generated using the model $y_{ij} = 5.0x_{ij} [\exp(\log^2(x_{ij}))]^{\gamma} u_i e_{ij}$. Here the individual effects e_{ij} and the area effects u_i were independently drawn as $\log(e_{ij}) \sim N(0, 1)$ and $\log(u_i) \sim N(0, 0.25)$ respectively, while the covariate values x_{ij} were drawn as $\log(x_{ij}) \sim N(3, 0.04)$. Five different values for the parameter γ (-1.0, -0.5, 0.0, 0.5, 1.0) were investigated, thus generating population data with different degrees of curvature. All other aspects of these simulations, including the estimators considered, were the same as in Set A. Table 3 presents results from this component of the simulation study.

Table 2

Average relative bias (ARB), average relative RMSE (ARRMSE), average coverage rate (ACR) and average interval width (AW) for model-based simulation Set A

Criterion	Estimator	Parameter Set					
		1	2	3	4	5	6
ARB, %	HJ-TrMBD	-82.68	-95.02	-98.08	-98.50	-98.29	-99.00
	HT-TrMBD	0.09	0.10	-0.14	-0.25	-0.03	0.04
	TrEP	0.08	0.09	-0.18	-0.48	-0.05	0.01
	HJ-LinMBD	12.01	4.09	-1.35	-5.54	-6.60	-9.88
	HT-LinEBLUP	13.39	5.18	-0.67	-5.24	-6.41	-9.67
ARRMSE	HJ-TrMBD	4.80	1.39	1.25	1.44	1.42	1.62
	HT-TrMBD	0.15	0.26	0.45	0.64	0.66	0.91
	TrEP	0.30	0.41	0.58	0.80	0.81	1.09
	HJ-LinMBD	1.11	1.41	1.85	1.99	2.06	2.69
	HT-LinEBLUP	0.79	0.54	0.64	0.92	0.93	1.31
ACR	HJ-TrMBD	0.99	0.98	0.97	0.95	0.94	0.92
	HT-TrMBD	0.94	0.91	0.89	0.89	0.89	0.88
	HJ-LinMBD	0.87	0.85	0.85	0.88	0.88	0.87
	HT-LinEBLUP	0.85	0.85	0.86	0.87	0.88	0.87
AW	HJ-TrMBD	1,592	22,688	140,452	52×10^4	35×10^5	44×10^6
	HT-TrMBD	219	4,414	34,105	14×10^4	11×10^5	15×10^6
	HJ-LinMBD	1,005	19,232	139,420	57×10^4	41×10^5	56×10^6
	HT-LinEBLUP	382	7,099	57,039	26×10^4	21×10^5	32×10^6

Table 3

Average relative bias (ARB), average relative RMSE (ARRMSE), average coverage rate (ACR) and average interval width (AW) for model-based simulation Set B

Criterion	Estimator	$\gamma = -1.0$	$\gamma = -0.5$	$\gamma = 0.0$	$\gamma = 0.5$	$\gamma = 1.0$
ARB, %	HT-TrMBD	4.92	0.66	0.14	-1.50	-8.75
	HJ-LinMBD	-0.21	0.04	0.12	0.16	-0.85
	HT-LinEBLUP	-0.19	0.04	0.13	0.17	-0.77
ARRMSE	HT-TrMBD	0.38	0.35	0.33	0.37	0.41
	HJ-LinMBD	0.56	0.36	0.34	0.53	1.20
	HT-LinEBLUP	0.38	0.30	0.29	0.36	0.56
ACR	HT-TrMBD	0.94	0.92	0.92	0.91	0.87
	HJ-LinMBD	0.91	0.92	0.92	0.92	0.90
	HT-LinEBLUP	0.93	0.94	0.94	0.93	0.92
AW	HT-TrMBD	0.04	2.50	211	29,070	5×10^6
	HJ-LinMBD	0.06	2.70	214	38,660	13×10^6
	HT-LinEBLUP	0.05	2.60	214	33,442	10×10^6

6.2 The design-based simulation study

This study used the same population and samples as the simulation studies described in Chandra and Chambers (2005) and Chambers and Tzavidis (2006), which was based on data obtained from a sample of 1,652 farms that participated in the Australian Agricultural and Grazing Industries Survey (AAGIS). A realistic population of 81,982 farms was defined by sampling with replacement from the original sample of 1,652 farms with probabilities proportional to their sample weights, all of which were strictly

greater than one. A total of 1,000 independent samples, each of size $n = 1,652$, were drawn from this fixed population by simple random sampling without replacement within strata defined by the 29 Australian agricultural regions represented in the AAGIS sample. These regions are the small areas of interest. Regional sample sizes were fixed to be the same as in this original sample, varying from a low of 6 to a high of 117, which allows an evaluation of the performance of the different estimation methods across a range of realistic small area sample sizes. Note that sampling fractions in these strata also varied disproportionately, ranging between 0.70

and 15.87 percent. The aim is to estimate average annual farm costs (TCC, measured in A\$) in each region using farm size (hectares) as the auxiliary variable. The same mixed model specification as in Chandra and Chambers (2005) is used. This includes an interaction term (zone by size) in the fixed effects and a random slope specification for the area effect. In its linear form the model does not fit the AAGIS sample data terribly well. This fit is improved (albeit marginally) when a log-scale linear specification is used. Our results are summarized in Table 4.

6.3 Discussion of simulation results

The most striking feature of Table 2 is the extremely large values of the averages relative bias of HJ-TrMBD under model-based model-calibrated weighting. The two best performers with respect to relative bias are HT-TrMBD, which is based on the same weights as HJ-TrMBD, and TrEP. An investigation of the reason for the poor performance of HJ-TrMBD revealed that summing the model-based model-calibrated weights (17) within small areas produced extremely variable estimates of the small area population sizes, implying that these weights cannot be considered as ‘multipurpose’ – they function well when used with variables that are reasonably correlated with the variable that defines the fitted value model, but can fail with other, less well correlated, variables (*e.g.*, the indicator variable for small area inclusion). We further note that this problem does not arise with the ‘standard’ empirical EBLUP weights for the total (6), as HJ-LinMBD performs consistently for all six of the scenarios explored in Set A of the simulation study. From now on we therefore focus our discussion on the four estimators, HT-TrMBD, TrEP, HJ-LinMBD and HT-LinEBLUP.

Table 2 shows that the average relative biases and the average relative RMSEs for HT-TrMBD are consistently lower than those generated by HJ-LinMBD and HT-LinEBLUP. The average relative biases of HT-TrMBD and TrEP are comparable. However, the average relative RMSEs of HT-TrMBD are consistently smaller than the TrEP. Furthermore, average coverage rates and interval widths for HT-TrMBD are better than those generated by HJ-LinMBD and HT-LinEBLUP. In comparison, for the same order of relative bias, the relative RMSEs of HT-LinEBLUP is smaller than that of HJ-LinMBD, and, although both estimators generate very similar coverage rates, confidence intervals generated via HT-LinEBLUP tend to have smaller average widths than those generated via HJ-LinMBD.

The plots in Figure 1 display the region-specific performance measures generated by these four estimators for the Set A simulations. These show that the relative bias and the relative RMSE values generated by HT-TrMBD are smaller than corresponding values for HJ-LinMBD and HT-LinEBLUP in all regions. With almost identical values of relative biases, the HT-TrMBD has smaller values of relative RMSEs than corresponding values for TrEP in all regions. Further, the relative bias and the relative RMSE of HJ-LinMBD and HT-LinEBLUP increase as the non-linearity in the data increases (*i.e.*, as we move from parameter set 1 to parameter set 6). We also see that HT-TrMBD generates better coverage rates across all regions compared with the coverage rates generated by HT-LinEBLUP and HJ-LinMBD.

Table 4
Average relative bias (ARB), average relative RMSE (ARRMSE) and average coverage rate (ACR) for design-based simulation using AAGIS data. Simulation standard errors of ARB and ARRMSE are shown in parentheses

Criterion	Estimator	Average of 29 regions	Average of 28 regions
ARB, %	HT-TrMBD	1.96 (0.20)	1.92 (0.11)
	HJ-LinMBD	-2.13 (0.15)	-2.21 (0.12)
	HT-LinEBLUP	2.98 (0.18)	3.36 (0.16)
	PseudoEBLUP	4.01 (0.22)	4.41 (0.20)
	JL	1.89 (0.19)	2.23 (0.17)
ARRMSE, %	HT-TrMBD	21.93 (4.47)	17.41 (1.18)
	HJ-LinMBD	20.15 (3.80)	16.91 (2.20)
	HT-LinEBLUP	19.87 (1.78)	19.30 (1.63)
	PseudoEBLUP	22.42 (2.52)	21.95 (2.46)
	JL	20.97 (1.48)	20.48 (1.31)
ACR	HT-TrMBD	0.89	0.92
	HJ-LinMBD	0.93	0.95
	HT-LinEBLUP	0.85	0.85

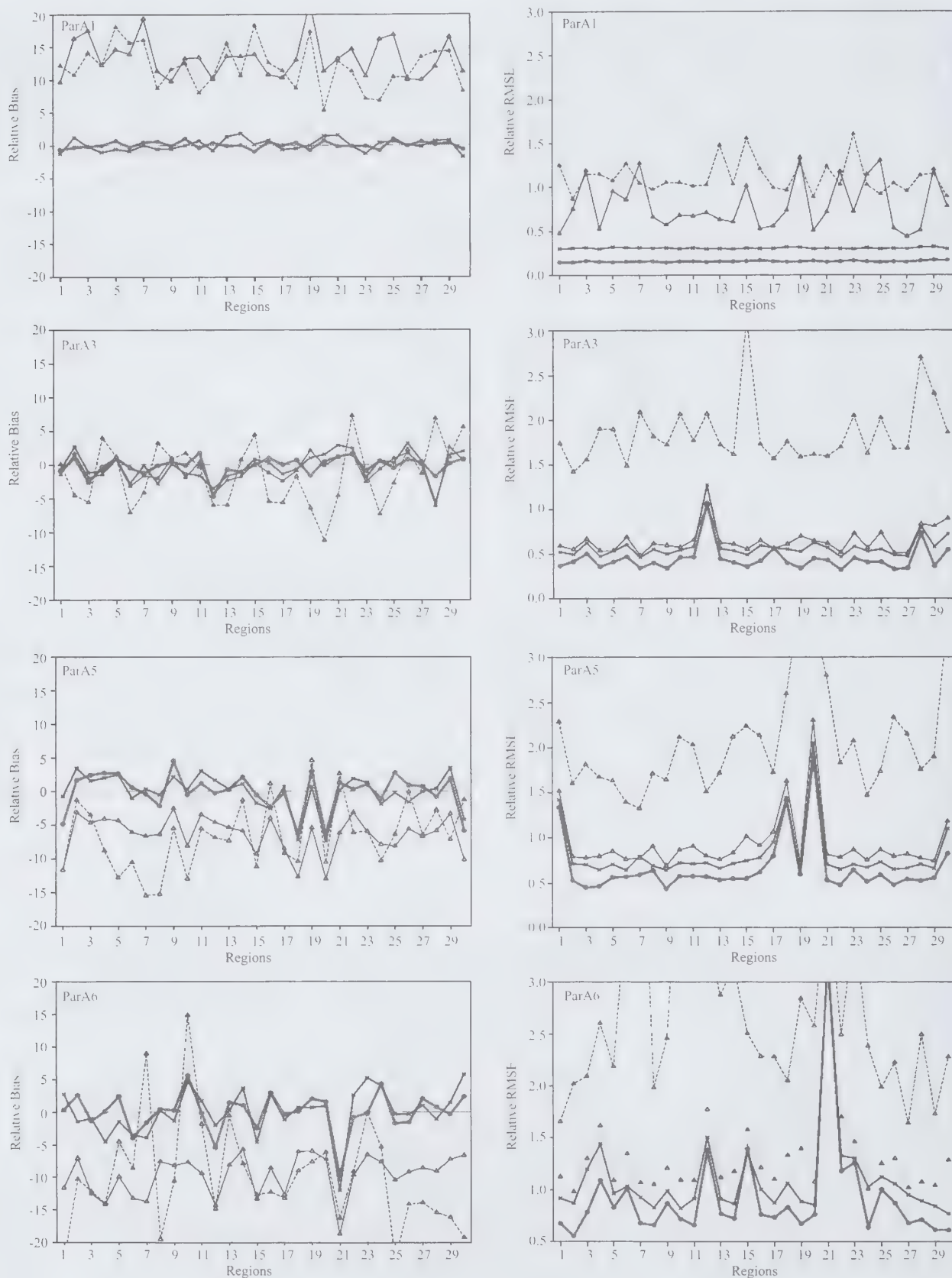


Figure 1 Area specific results for HT-TrMBD (solid line, ●), TrEP (thick line, ×), HT-LinEBLUP (thin line Δ) and HJ-LinMBD (dashed line, Δ) under parameter sets 1 (ParA1), 3 (ParA3), 5 (ParA5) and 6 (ParA6). Left column is Relative Bias (%) and right column is Relative RMSE (%)

Overall, these results show that when the model for the underlying population is non-linear there can be significant gains from the use of HT-type MBD estimators for small area means based on the model-calibrated weights (17) compared with standard linear mixed model-based estimators like HJ-LinMBD and HT-LinEBLUP. They also show that the indirect estimator HT-LinEBLUP performs relatively better than the direct estimator HJ-LinMBD in these situations. The indirect predictor TrEP based on log-transformed model (9) performs well in terms of relative bias but is less efficient than the MBD estimator under the same model.

In Set B of the model-based simulations we investigated the robustness of model-based model-calibrated direct estimation to misspecification of the non-linear model. The results in Table 3 show that in this case the biases generated by HT-TrMBD increase as the actual non-linear model deviates more from the assumed non-linear model ($\gamma = 0.0$ in the table). However, these biases are offset by small variability, so in terms of average relative RMSE, HT-TrMBD still performs as well or better than HT-LinEBLUP and continues to dominate HJ-LinMBD. The biases generated by HJ-LinMBD and HT-LinEBLUP are of the same order, while the average relative RMSE of HT-LinEBLUP dominates that of HJ-LinMBD. Average coverage rates for HT-LinEBLUP are marginally better than those of HJ-LinMBD and HT-TrMBD, but the average widths of the confidence intervals underpinning these rates tended to be smallest for HT-TrMBD, followed by HT-LinEBLUP and then HJ-LinMBD. Overall, our model-based simulation results for Set B indicate that although MBD-based SAE with model-based model-calibrated weights is susceptible to model misspecification bias, the overall performance of this approach appears relatively unaffected by slight deviations from the assumed non-linear model.

In Table 4 and Figure 2 we present the average and region-specific performance measure generated by different SAE methods for AAGIS data respectively. These results show that the average relative bias of HT-TrMBD is smaller than that of both HT-LinEBLUP and HJ-LinMBD, while the average relative RMSE of HT-TrMBD is marginally larger than the corresponding values for HJ-LinMBD and HT-LinEBLUP. Inspection of Figure 2 shows that this result is essentially due to one region (21) in the original AAGIS sample that contained a massive outlier (TCC > A\$30,000,000). This outlier was included in the simulation population (twice) and then selected (in one case, twice) in 37 of the 1000 simulation samples, leading to completely unrealistic estimates for region 21 being generated by HT-TrMBD and HJ-LinMBD. The right-hand column in Table 4 therefore shows the average performances of the different

methods when this region is excluded. Here we see that now HT-TrMBD and HJ-LinMBD are essentially on a par, with both dominating HT-LinEBLUP. The fact that HT-TrMBD does not provide significant gains over HJ-LinMBD in this case reflects the fact that the raw-scale and log-scale linear mixed models used in these estimators both provide relatively poor fits to the AAGIS data.

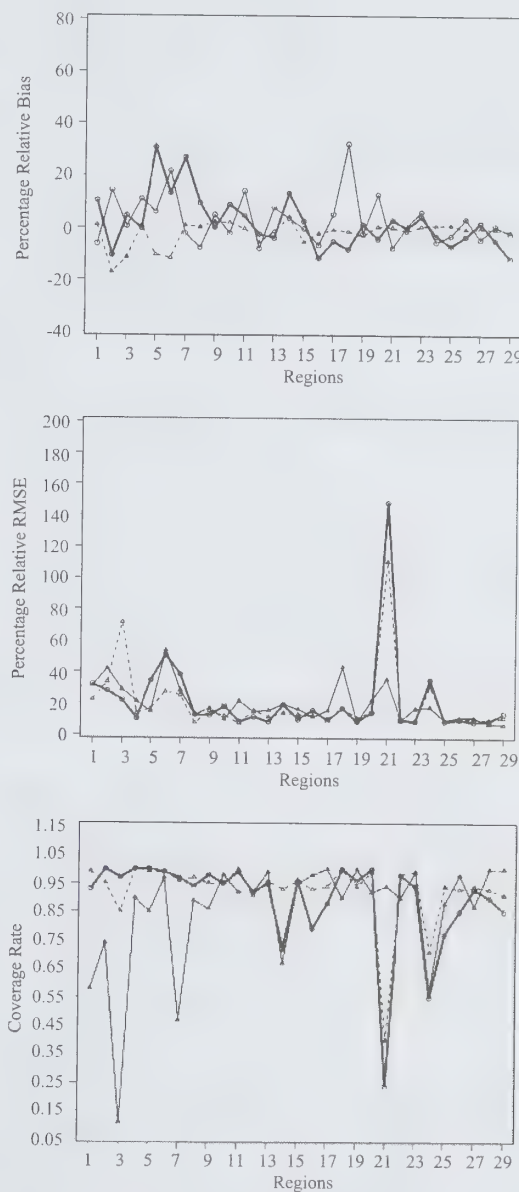


Figure 2 Region-specific simulation results for HT-TrMBD (thick line, \circ), HT-LinEBLUP (thin line Δ) and HJ-LinMBD (dashed line, Δ) in design-based simulations based on the AAGIS data. Plots show (in order from the top), RB (%), RRMSE (%) and CR. Regions are ordered in terms of increasing population size

7. Conclusions and further research

The simulation results discussed in the previous section show that combining model-based model-calibrated weights with direct estimation can bring significant gains in SAE efficiency if the population data are clearly non-linear. As one would expect, these gains are less when the assumed non-linear model is misspecified. Although we do not provide the details, our conclusions were essentially unaffected when we carried out similar simulations using gamma distributed random effects.

Our main caveat concerning the use of the model-based model-calibrated weights (17) for SAE is their specificity. These weights do not appear to have the same ‘multi-purpose’ characteristics as standard EBLUP weights for the total based on linear mixed models. Further research is therefore required on how to build model-calibrated weights for SAE that are more ‘general purpose’. It is to be expected that such weights would not be as efficient as the variable specific weights (17), but hopefully this will be more than offset by their increased utility. A further issue that is extremely important in practice is that positively skewed survey variables can also take zero (or even negative) values. For example, economic variables like debt and capital expenditure often take zero values, while variables defined as the difference of two non-negative quantities (e.g., profit, which is the difference between income and expenditure) can be negative. Karlberg (2000b) uses a mixture model to characterise data that are a mix of zeros and strictly positive values. This type of model can be used in model-based model-calibrated weighting.

Finally, we note that using a transformation-based MBD approach where the usual linear model assumptions are only approximately valid (the situation considered in this paper) is not the only approach that has been suggested for this problem. Two alternative approaches in the literature are the pseudo-EBLUP (Rao 2003, section 7.2.7) and the model-assisted EB-type estimator of Jiang and Lahiri (2006). Recollect from (8) that the EBLUP is defined by replacing the unknown area i mean m_{iy} by an estimate of its expected value given the observed sample values of Y in area i and the area i values of \mathbf{X} . Let π_{ij} denote the sample inclusion probability of population unit j in small area i . The pseudo-EBLUP is then defined by replacing m_{iy} by an estimate of its expected value given the value of its design-consistent estimate

$$\hat{m}_{iy}^{\pi} = \left(\sum_{j \in s_i} \pi_{ij}^{-1} \right)^{-1} \sum_{j \in s_i} \pi_{ij}^{-1} y_{ij} = \sum_{j \in s_i} \tilde{w}_{ij} y_{ij} \quad (21)$$

and the area i values of \mathbf{X} . That is, under (3) the pseudo-EBLUP of m_{iy} is

$$\hat{m}_{iy}^{\text{psuedoEBLUP}}$$

$$\begin{aligned} &= \hat{E}\{m_{iy} | \hat{m}_{iy}^{\pi}, \mathbf{x}_{is}, \mathbf{x}_{ir}\} \\ &= \bar{\mathbf{x}}_i' \hat{\beta}_{\bar{w}} + (\bar{\mathbf{g}}_i' \hat{\Sigma}_{u\bar{w}} \bar{\mathbf{g}}_{i\bar{w}}) \\ &\quad \left(\bar{\mathbf{g}}_{i\bar{w}}' \hat{\Sigma}_{u\bar{w}} \bar{\mathbf{g}}_{i\bar{w}} + \hat{\sigma}_{e\bar{w}}^2 \sum_{j \in s_i} \tilde{w}_{ij}^2 \right)^{-1} (\hat{m}_{iy}^{\pi} - \bar{\mathbf{x}}_{i\bar{w}}' \hat{\beta}_{\bar{w}}) \end{aligned} \quad (22)$$

where $\hat{\beta}_{\bar{w}}$, $\hat{\Sigma}_{u\bar{w}}$ and $\hat{\sigma}_{e\bar{w}}^2$ are pseudo-maximum likelihood estimates based on the weights \tilde{w}_{ij} and $\bar{\mathbf{g}}_{i\bar{w}}$ and $\bar{\mathbf{x}}_{i\bar{w}}$ are design-consistent estimates of $\bar{\mathbf{g}}_i$ and $\bar{\mathbf{x}}_i$ that are defined in exactly the same way as \hat{m}_{iy}^{π} above. Under the same model the Jiang and Lahiri (2006) model-assisted EB-type approach leads to an estimator that is also defined by conditioning on the value of \hat{m}_{iy}^{π} ,

$$\begin{aligned} \hat{m}_{iy}^{JL} &= \sum_{j \in s_i} \tilde{w}_{ij} \hat{E}\{\hat{E}(y_{ij} | \mathbf{x}_{ij}, \mathbf{u}_i) | \hat{m}_{iy}^{\pi}, \mathbf{x}_i\} \\ &= \bar{\mathbf{x}}_{i\bar{w}}' \hat{\beta} + \{\bar{\mathbf{w}}_{is}' (\mathbf{g}_{is}' \hat{\Sigma}_u \mathbf{g}_{is} + \hat{\sigma}_e^2 \mathbf{I}_{is}) \bar{\mathbf{w}}_{is}\}^{-1} \\ &\quad \{\bar{\mathbf{w}}_{is}' \mathbf{g}_{is}' \hat{\Sigma}_u \mathbf{g}_{is}' \bar{\mathbf{w}}_{is}\} (\hat{m}_{iy}^{\pi} - \bar{\mathbf{x}}_{i\bar{w}}' \hat{\beta}) \end{aligned} \quad (23)$$

where $\bar{\mathbf{w}}_{is}$ is the vector of standardised sample weights \tilde{w}_{ij} in area i . Note that in (23) we use optimal (i.e., ML or REML) estimates for model parameters.

Both (22) and (23) are essentially motivated by the idea of estimating the area i mean by its conditional expectation under (3) given the value of the usual design-consistent estimator (21) for this quantity. As such, they are indirect estimators like the HT-LinEBLUP. Under (3), neither will be as efficient as the HT-LinEBLUP, while if (9) rather than (3) holds, then both estimators rely on the design consistency of \hat{m}_{iy}^{π} for robustness. Since relying on a large sample property of a small sample statistic seems rather optimistic, we prefer to tackle the model specification problem directly, replacing (3) by (9) and using the transformation-based MBD approach described in section 5. Values of average relative bias and average relative RMSE for the pseudo-EBLUP (22) and the Jiang and Lahiri estimator (23) are shown in Table 4. It is interesting to note that neither estimator appears to perform any better than the standard EBLUP in these design-based simulations, and all three are substantially outperformed in terms of average relative RMSE by the two MBD-type estimators that were investigated in this study. Clearly the results of a single (but reasonably realistic) simulation study should not be considered as anything more than indicative. However, they do provide some evidence that asymptotic design-based properties are no guarantee of small area estimation performance.

The indirect predictor (20) of the small area mean is obtained by using well known prediction-based ideas. Under log transformed models, there are alternative approaches to obtain better indirect predictor for small area mean. For example, Slud and Maiti (2006) described an

indirect predictor for the small area mean under an area level version of the log transformed model (9). Berg (2009, private communication) follows the Slud-Maiti approach to obtain a predictor for small area mean under a random intercepts specification of the unit level log transformed model (9). However, like the Slud-Maiti predictor, Berg's predictor ignores the bias correction necessary after back-transformation to the raw scale. The empirical properties of this predictor have yet to be examined.

Acknowledgements

The first author gratefully acknowledges the financial support provided by a PhD scholarship from the U.K. Commonwealth Scholarship Commission. Constructive comments from Editor, Associate Editor and two referees are also gratefully acknowledged. They resulted in the revised version of the article representing a considerable improvement on the original.

References

- Bates, D.M., and Pinheiro, J.-C. (1998). Computational Methods for Multilevel Models. <http://franz.stat.wisc.edu/pub/NLME/>.
- Carroll, R., and Ruppert, D. (1988). *Transformation and Weighting in Regression*. New York: Chapman and Hall.
- Chambers, R., and Tzavidis, N. (2006). M-quantile models for small area estimation. *Biometrika*, 93, 255-268.
- Chandra, H., and Chambers, R.L. (2005). Comparing EBLUP and C-EBLUP for small area estimation. *Statistics in Transition*, 7, 637-648.
- Chandra, H., and Chambers, R. (2009). Multipurpose weighting for small area estimation. *Journal of Official Statistics*, 25(3), 379-395.
- Chandra, H., Salvati, N. and Chambers, R. (2007) Small area estimation for spatially correlated populations. A comparison of direct and indirect model-based methods. *Statistics in Transition*, 8, 887-906.
- Chen, G., and Chen, J. (1996). A transformation method for finite population sampling calibrated with empirical likelihood. *Survey Methodology*, 22, 139-146.
- Harville, D.A. (1977). Maximum likelihood approaches to variance component estimation and to related problems. *Journal of the American Statistical Association*, 72, 320-338.
- Hidiroglou, M.A., and Smith, P.A. (2005). Developing small area estimates for business surveys at the ONS. *Statistics in Transition*, 7, 527-539.
- Jiang, J., and Lahiri, P. (2006). Estimation of finite population domain means: A model-assisted empirical best prediction approach. *Journal of the American Statistical Association*, 101, 301-311.
- Karlberg, F. (2000a). Population total prediction under a lognormal superpopulation model. *Metron*, LVIII, 53-80.
- Karlberg, F. (2000b). Survey estimation for highly skewed populations in the presence of zeroes. *Journal of Official Statistics*, 16, 229-241.
- Longford, N.T. (2007). On standard errors of model-based small-area estimators. *Survey Methodology*, 33, 69-79.
- McCulloch, C.E., and Searle, S.R. (2001). *Generalized, Linear and Mixed Models*. New York: John Wiley & Sons, Inc.
- Prasad, N.G.N., and Rao, J.N.K. (1990). The estimation of the mean squared error of small area estimators. *Journal of the American Statistical Association*, 85, 163-171.
- Rao, J.N.K. (2003). *Small Area Estimation*. New York: John Wiley & Sons, Inc.
- Royall, R.M. (1976). The linear least-squares prediction approach to two-stage sampling. *Journal of the American Statistical Association*, 71, 657-664.
- Royall, R.M., and Cumberland, W.G. (1978). Variance estimation in finite population sampling. *Journal of the American Statistical Association*, 73, 351-358.
- Slud, E. V., and Maiti, T. (2006). Mean squared error estimation in transformed Fay-Herriot models. *Journal of the Royal Statistical Society, Series B*, 68(2), 239-257.
- Tzavidis, N., Salvati, N., Pratesi, M. and Chambers, R. (2008). M-quantile models with application to poverty mapping. *Statistical Methods And Applications*, 17, 393-411.
- Valliant, R., Dorfman, A.H. and Royall, R.M. (2000). *Finite Population Sampling and Inference*. New York: John Wiley & Sons, Inc.
- Wu, C., and Sitter, R.R. (2001). A model calibration approach to using complete auxiliary information from survey data. *Journal of the American Statistical Association*, 96, 185-193.

The construction of stratified designs in R with the package *stratification*

Sophie Baillargeon and Louis-Paul Rivest¹

Abstract

This paper introduces a R-package for the stratification of a survey population using a univariate stratification variable X and for the calculation of stratum sample sizes. Non iterative methods such as the cumulative root frequency method and the geometric stratum boundaries are implemented. Optimal designs, with stratum boundaries that minimize either the CV of the simple expansion estimator for a fixed sample size n or the n value for a fixed CV can be constructed. Two iterative algorithms are available to find the optimal stratum boundaries. The design can feature a user defined certainty stratum where all the units are sampled. Take-all and take-none strata can be included in the stratified design as they might lead to smaller sample sizes. The sample size calculations are based on the anticipated moments of the survey variable Y , given the stratification variable X . The package handles conditional distributions of Y given X that are either a heteroscedastic linear model, or a log-linear model. Stratum specific non-response can be accounted for in the design construction and in the sample size calculations.

Key Words: Linear models; Log-linear models; Optimal stratification; Survey sampling; Take-all stratum; Take-none stratum.

1. Introduction

The establishment of strata and the planning of a stratified design have been important topics in survey sampling, since the pioneering contributions of Dalenius more than sixty years ago. This work is concerned with univariate stratification where the strata are constructed using a positive stratification variable X known for all the units of the population. X is assumed to be related to the survey variable Y . Stratum h contains all the units with an X -value in the interval $[b_{h-1}, b_h)$ for $h = 1, \dots, L$ such that $b_0 = \min X$ and $b_L = \max X + 1$, where $\min X$ and $\max X$ are respectively the minimum and the maximum values of the stratification variable.

The determination of optimal stratum boundaries has a long history, see chapter 5A of Cochran (1977). The cumulative root frequency method ($\text{cum}\sqrt{f}$) of Dalenius and Hodges (1959) provides an approximate solution to this problem. Instances where X has a skewed distribution are frequent in business surveys and have been given a special emphasis. Gunning and Horgan (2004) proposed a geometric stratification method and Hidiroglou (1986) argued that the large units should be put in a take-all stratum. Rather than relying on an approximate method for constructing the strata, Lavallée and Hidiroglou (1988) suggested an iterative algorithm that gives the optimal boundaries for a particular X variable. Their algorithm sometimes fails to converge (Detlefsen and Veum 1991) and Slanta and Krenzke (1996) have shown that in some cases the optimal boundaries are not uniquely defined. Alternative methods, such as the search algorithm of Kozak (2004), have been

proposed to alleviate some of these difficulties. The assumption that the survey variable Y is the same as the stratification variable X is not realistic when calculating sample sizes and several authors, including Dayal (1985) and Sigman and Monsour (1995), proposed to allocate the sample to the strata on the basis of the *anticipated* moments of Y knowing that X is in $[b_{h-1}, b_h)$. Sweet and Sigman (1995) and Rivest (1999, 2002) suggested using these anticipated moments in the stratification algorithm of Lavallée and Hidiroglou (1988). Recently, Baillargeon and Rivest (2009) showed that putting the small units in a take-none stratum, which is not sampled, might reduce the sample size needed to reach a predetermined precision level.

This article introduces the R-package *stratification* that implements most of the methods presented above. It provides a friendly computer environment to build stratified designs and to evaluate their performance on some real populations. This package is presented by revisiting examples in the stratification literature selected to illustrate its important features. The four functions of *stratification* with the prefix *strata* construct stratified sampling designs. These functions are *strata.cumrootf*, *strata.geo*, *strata.LH*, and *strata.bh*. The first two implement the simple $\text{cum}\sqrt{f}$ and geometric stratification methods. The function *strata.LH* derives optimal stratified sampling plans using iterative algorithms while the last function handles user defined stratum boundaries. These four functions construct strata, determine stratum sample sizes and calculate the precision of the simple expansion estimator \bar{y}_s of \bar{Y} , the population mean of some survey variable Y related to the stratification variable X .

1. Sophie Baillargeon, Département de mathématiques et de statistique, 1045, avenue de la médecine, Université Laval, Québec, (Qc) Canada G1V 0A6. E-mail: sophie.baillargeon@mat.ulaval.ca; Louis-Paul Rivest, Département de mathématiques et de statistique, 1045, avenue de la médecine, Université Laval, Québec, (Qc) Canada G1V 0A6. E-mail: louis-paul.rivest@mat.ulaval.ca.

The four *strata*-functions use Hidioglou and Srinath's (1993) rule to allocate the n units in the sample to the strata. The stratum sample sizes are proportional to $N_h^{2q_1} \bar{Y}_h^{2q_2} S_{yh}^{2q_3}$, where N_h is the size of stratum h , and \bar{Y}_h and S_{yh}^2 are the anticipated mean and variance of Y in stratum h . In the *strata*-functions, an allocation rule is specified by the argument *alloc* that contains the exponents (q_1, q_2, q_3) ; Neyman's allocation corresponds to *alloc*= $c(1/2, 0, 1/2)$. A *strata*-function takes as an input the population vector of the stratification variable X , the number of strata L_s , and a total sample size n or a target CV for the simple expansion estimator \bar{y}_s . Its output is an R-object of class "strata" that defines a stratified design. It contains a set of strata determined by their upper boundaries $\{b_h\}$ and stratum population and sample sizes, N_h and n_h . There is a fifth function in *stratification* called *var.strata* that takes as an input an R-object of class strata and a population vector of a survey variable Y and returns the variance of \bar{y}_s for the input variable Y and the input stratified design.

The text contains R instructions to be typed in an R command window; these lines start with $>$. It also presents outputs printed in an R command window. A special typeface allows an easy identification of these R instructions and print-outs in the text. The appendix contains a summary table that lists all the possible arguments of the five *stratification* functions. When using this package, the R-instruction *help(stratification)* calls a clickable help file that provides detailed information on the package and examples that can be pasted in a command window.

2. Basic stratification methods

This section discusses two elementary stratification methods, the cumulative root frequency method of Dalenius and Hodges (1959) and the geometric method of Gunning and Horgan (2004). These two methods are exact; they do not rely on an iterative algorithm. Throughout this section $Y = X$, so that the variance of \bar{y}_s is evaluated using the values of the stratification variable X . Using the same variable to stratify a population and to evaluate the precision of survey estimates might underestimate their variances. The calculation of variances when $Y \neq X$ is considered in Section 4.

2.1 Cumulative root frequency method

This stratification algorithm, presented in chapter 5A of Cochran (1977), is implemented by the function *strata.cumrootf*. Its arguments are *x*, the population vector of the stratification variable, *nclass* the number of bins of equal size for the *x*-variable, a target CV for \bar{y}_s or a predetermined sample size *n*, the number of strata L_s , and an allocation rule *alloc*. This algorithm pools the *nclass* bins into L_s strata in such a way that the sums of the square

roots of the bin frequencies are approximately equal for the L_s strata.

As an illustration, consider the proportion of industrial loans of $N = 13,435$ banks used in Cochran (1961). We stratify this population and evaluate the sample size needed for \bar{y}_s to have a CV of 5% when Neyman allocation is used. The following R-code creates the vector of the stratification variable *loans* from Table 2 of McEvoy (1956). The function *strata.cumrootf* is then applied to the *loans* variable. Following Table 2 of Cochran (1961), *nclass* is set to 20 so that the strata will be created using 20 bins and $L_s=3$ strata will be constructed. The output is placed in *cum*, an R-object of class strata. Typing *cum* or *print(cum)* in the R command window prints details of the sampling plan. The input arguments, either the default or as specified by the user, appear first. Then stratum information is provided such as boundaries, sizes N_h and sample sizes n_h . The third part of the print-out provides information about the sampling properties of \bar{y}_s .

```
> values <- c(seq(0.5, 9.5, 1), seq(12.5, 97.5, 5))
> nrep <- c(1985, 261, 339, 405, 474, 478, 506, 569, 464, 499,
  2157, 1581, 1142, 746, 512, 376, 265, 207, 126, 107, 82, 50,
  33, 25, 16, 14, 2, 3)
> loans <- rep(values, nrep)
> cum <- strata.cumrootf(x = loans, nclass = 20, CV = 0.05,
  Ls = 3, alloc = c(0.5, 0, 0.5))
> cum
```

Given arguments:

```
x = loans
nclass = 20, CV = 0.05, Ls = 3
allocation : q1 = 0.5, q2 = 0, q3 = 0.5
model = none
```

Strata information:

	rh	bh	anticip.Mean	anticip.var	Nh	nh	fh
Stratum 1	1	10.2	4.12	10.46	5980	14	0.00
Stratum 2	1	29.6	17.92	27.74	5626	20	0.00
Stratum 3	1	98.5	44.47	165.83	1829	16	0.01
Total					13435	50	0.00

Total sample size: 50

Anticipated population mean: 15.39408

Anticipated CV: 0.0494897

In the Given arguments, *model=none* means that the sampling properties of \bar{y}_s , presented at the end of the print-out, are evaluated at $Y = X$, that is for the *loans* variable. Its mean is 15.39408 and the anticipated CV of 0.0494897 is that of the estimator \bar{y}_s of the mean of the variable *loans* obtained with this sampling design. The stratum boundaries given in this output are (10.2, 29.6, 98.5), they are equal to those appearing at the bottom of page 349 of Cochran (1961), once the rounding used for creating the vector *loans* is accounted for. In the Strata Information, r_h refers to the stratum response rates that are discussed in Section 5.1. The R-object *cum* contains several elements that are listed by the command *names(cum)*.

```
> names(cum)
[1] "Nh" "nh" "n" "nh.nonint" "certain.info"
[6] "opti.criteria" "bh" "meanh" "varh" "mean"
[11] "stderr" "CV" "stratumID" "nclassh" "takeall"
[16] "call" "date" "args"
```

An element in the `cum` strata object can be printed by typing `cum$` followed by the name of the object. For instance the `cum$stratumID` prints the stratum of each unit in the population. The variable `cum$nclash` is specific to the `strata.cumrootf` function; it gives how the `nclash=20` original bins have been pooled into three strata;

```
> cum$nclash
[1] 2 4 14
```

Thus, in this stratification, strata 1, 2 and 3 contain respectively 2, 4 and 14 of the `nclash=20` original bins.

2.2 Geometric method

The geometric stratification method has been introduced by Gunning and Horgan (2004). It sets the stratum boundaries to $b_h = \min X \times (\max X / \min X)^{h/L}$, for $h = 1, \dots, L - 1$. Once the boundaries b_h are determined, the stratum sample size calculations are the same as those carried out in `strata.cumrootf`.

As an illustration we stratify the four populations presented in Gunning et Horgan (2004), Debtors, USbanks, UScities, and UScolleges, into `Ls=5` strata. The last three populations were considered in Cochran's (1961) investigations. These four populations are stored in *stratification*; the command `data(Debtors)` calls the first one. Rather than specifying a target CV we set the sample size to $n = 100$ following Gunning and Horgan (2004). The following commands create the R-object `pop1` that contains the stratified design for the Debtors population.

```
> data(Debtors)
> pop1 <- strata.geo(x = Debtors, n = 100, Ls = 5,
  alloc = c(0.5, 0, 0.5))
```

Table 1 summarizes the geometric stratified designs for the four study populations. It reproduces Table 4 of Gunning and Horgan (2004) partially. There are however some minor differences caused by different rounding strategies. More details about *stratification* rounding methods are available in the help file.

Table 1
Stratified designs for four populations with $n = 100$

Population	CV		1	2	3	4	5
Debtors	0.0359	b_h	148.28	549.67	2,037.60	7,553.33	
		N_h	1,054	1,267	732	265	51
		n_h	3	14	27	33	23
UScities	0.0145	b_h	18.17	33.01	59.98	108.98	
		N_h	364	418	130	87	39
		n_h	18	28	17	20	17
UScolleges	0.0183	b_h	434.00	941.76	2,043.61	4,434.60	
		N_h	94	255	198	74	56
		n_h	3	15	27	20	35
USbanks	0.0107	b_h	118.59	200.92	340.39	576.68	
		N_h	114	116	64	39	24
		n_h	13	20	25	18	24

2.3 Take-all stratum

In Table 1, the fifth stratum for the USbanks population is a take-all stratum since $n_5 = N_5 = 24$. Under Neymann allocation, the fifth stratum gets a sample size n_5 larger than the stratum size N_5 . Then `strata.geo` automatically identifies this stratum as a take-all stratum and allocates the $n - N_5$ units for the first four strata using Neyman allocation. This adjustment is important to have a sample size of $n = 100$ as specified in the `strata.geo` arguments.

To illustrate this point, we use the function `strata.bh` to make an allocation without a take-all stratum adjustment. This function allocates the sample and calculates the precision of \bar{y}_s for a predetermined set of stratum boundaries. By setting `takeall.adjust=FALSE`, Neyman allocation is used in the five strata and since $n_5 > N_5$ one has $n_5 = N_5$. The following R-code gets the geometric stratum boundaries $\{b_h\}$ in the strata object `adjust`; it then uses the `strata.bh` function with the geometric stratum boundaries to get the sampling design without adjusting for a take-all stratum five in the `noadjust` strata object.

```
> data(USbanks)
> adjust <- strata.geo(x = USbanks, n = 100, Ls = 5,
  alloc = c(0.5, 0, 0.5))
> noadjust <- strata.bh(x = USbanks, bh = adjust$bh,
  n = 100, Ls = 5, alloc = c(0.5, 0, 0.5), takeall = 0,
  takeall.adjust = FALSE)
```

The two designs are presented in Table 2. Failing to include a take-all stratum yields a sample size of $n = 99$, smaller than the target $n = 100$. In this case, the unrounded sample size for stratum 5 is `noadjust$nh.noint[5]=25.40` for $N_5 = 24$ units. Note that when n is large or when the target CV is small, it is possible to get several take-all strata.

Table 2
Stratified designs obtained with and without an automatic adjustment for a take-all stratum

	n		1	2	3	4	5
		b_h	118.59	200.92	340.39	576.68	
		N_h	114	116	64	39	24
adjust	100	n_h	13	20	25	18	24
noadjust	99	n_h	13	20	24	18	24

2.4 Adding a take-all stratum

We now consider the data base on $N = 284$ Swedish municipalities given in the appendix of Särndal, Swensson and Wretman (1992). The following instructions use the geometric method to stratify this population in `Ls=5` strata using the variable `REV84`, the 1984 real estate values. The power allocation with exponent 0.7 and `alloc=c(0.35,0.35,0)` is used. The R-object of class `strata geo` contains the stratified design. The command `plot(geo)` produces the plot presented in Figure 1. It provides a histogram of the

stratification variable with the stratum boundaries and a summary table for the stratified design.

```
> data(Sweden)
> geo <- strata.geo(x = Sweden$REV84, CV = 0.05, Ls = 5,
  alloc = c(0.35, 0.35, 0))
```

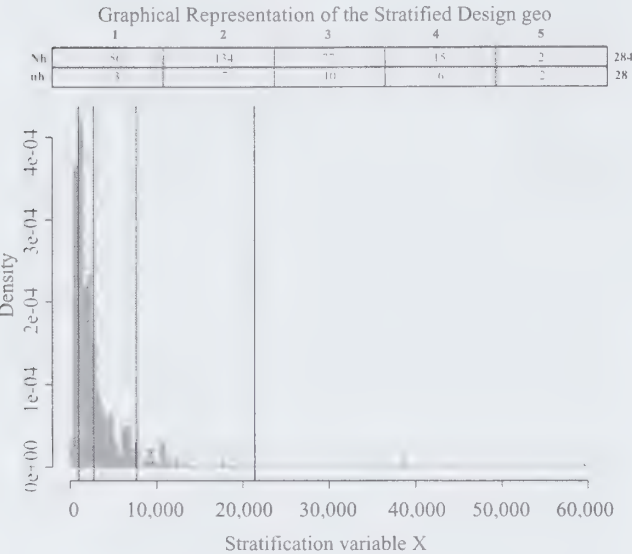


Figure 1 Plot of the R-object geo

Figure 1 shows that the geometric stratification method puts two of the three extreme REV84 values in a take-all stratum. The following Rcode creates cum a stratified design for this population using the cum \sqrt{f} method. The application of this stratification method is awkward since the bins have length $\{\max(\text{REV84}) - \min(\text{REV84})\} / 50 = 1191$. Considering Figure 1 most of the bins have a null frequency; indeed stratum 5 comprises 43 of the 50 bins. This design does not have a take-all stratum. To calculate the sample sizes obtained by requesting a take-all stratum one can use the function `strata.bh`, with the cum \sqrt{f} boundaries stored in `cum3$bh`, with the command `takeall=1`. This gives the third sampling plan in Table 3. The fourth sampling plan of Table 3 `cum3` is created by setting the sample size in stratum 5 of the cum \sqrt{f} design equal to its population size with the command `cum3$nh[5]<-cum3$Nh[5]`. The variance of the estimate \bar{y}_s for the variable REV84 using this fourth sampling design is calculated using `var.strata`.

```
> cum <- strata.cumrootf(x = Sweden$REV84, nclass = 50,
  CV = 0.05, Ls = 5, alloc = c(0.35, 0.35, 0))
> cum2 <- strata.bh(x = Sweden$REV84, bh = cum$bh, CV = 0.05,
  Ls = 5, takeall = 1, alloc = c(0.35, 0.35, 0))
> cum3 <- cum
> cum3$nh[5] <- cum3$Nh[5]
> cum3.var <- var.strata(cum3, y = Sweden$REV84)
```

Table 3
Four stratified designs for the population of Swedish municipalities

Method		1	2	3	4	5	n	CV
geometric	N_h	56	134	77	15	2		
	n_h	3	7	10	6	2	28	4.83
cum \sqrt{f}	N_h	120	70	52	27	15		
	n_h	7	7	9	8	10	41	4.87
	n_h^{modif1}	2	2	3	2	15	24	4.44
	n_h^{modif2}	7	7	9	8	15	46	2.29

Table 3 highlights that the sampling fraction in the fifth stratum drives the value of n . The cum \sqrt{f} design appears to be less efficient than the geometric design since its sampling fraction in stratum 5 is $10/15 = 67\%$. Requesting a take-all stratum gives a value of n comparable to that obtained with the geometric design. The REV84 population has three outliers that were identified in Table 1. The geometric and cum \sqrt{f} stratification methods depend heavily on the maximum X -value; therefore before applying these techniques it might be wise to put the three outliers aside. This is considered in the next section.

The simple *ad hoc* method to arbitrarily change the stratum sample sizes presented in this section can be applied in several situations. For instance, when some strata have samples of size 1, they can be increased to 2 in order to have an unbiased variance estimator.

2.5 Certainty stratum

In a stratified design it might be useful to constrain some units to be sampled, before constructing the strata. The argument `certain` available in the four `strata`-function makes this possible. As an example we revisit the comparison of the cum \sqrt{f} and the geometric sampling designs presented in Table 3. The three large municipalities highlighted in Figure 1 are put in a certainty stratum, and the $N = 281$ remaining municipalities are stratified into $Ls=4$ strata using the two stratification methods. The R-code for constructing these two designs is given below. The command `x=sort(Sweden$REV84)` orders the municipalities by increasing REV84; thus the three large municipalities are entries 282, 283 and 284 of the sorted vector. The two R objects of class `strata`, `geo_cer` and `cum_cer`, each contain an element `certain.info` that provides information on the certainty stratum.

```
> geo_cer <- strata.geo(x = sort(Sweden$REV84), CV = 0.05,
  Ls = 4, alloc = c(0.35, 0.35, 0), certain = 282:284)
> cum_cer <- strata.cumrootf(x = sort(Sweden$REV84),
  nclass = 50, CV = 0.05, Ls = 4, alloc = c(0.35, 0.35, 0),
  certain = 282:284)
> cum_cer$certain.info

  Nc      meanc
3.00  38923.67
```

In Table 4, the $\text{cum}\sqrt{f}$ design is more efficient than the geometric design. Putting the three large municipalities in a certainty stratum is helpful since the sample sizes in Table 4 are smaller than those of Table 3. The argument can force any set of units in the sample. It can be used to include units that are extreme for a secondary variable, different from the stratification variable, or that have a history of high volatility.

Table 4
Two stratified designs for the Swedish municipalities constructed with a certainty stratum

Method		1	2	3	4	5	<i>n</i>	CV
geometric	N_h	42	116	88	35	3	24	4.71
	n_h	2	5	7	7	3		
$\text{cum}\sqrt{f}$	N_h	127	79	46	29	3	19	4.72
	n_h	3	4	4	5	3		

3. Optimization method

The stratification methods introduced in Section 2 do not always give an optimal stratified design, that minimizes the sample size n needed to reach the target CV (or minimizes the CV for a fixed n). This section introduces the function `strata.LH` that allows the determination of optimal designs. The name LH stands for Lavallée and Hidiroglou (1988) who pioneered the construction of optimal stratified designs for real life survey populations. In a stratified design with a take-all stratum, the variance of the simple expansion estimator is given by

$$\text{Var}(\bar{Y}_s) = \sum_{h=1}^{L-1} \left(\frac{N_h}{N} \right)^2 \left(\frac{1}{(n - N_L)a_h} - \frac{1}{N_h} \right) S_{yh}^2,$$

where $\{a_h\}$ is the allocation rule for setting stratum sample sizes. The n that ensures a CV of c is given by

$$n = N_L + \frac{\sum_{h=1}^{L-1} N_h^2 S_{yh}^2 / (a_h N^2)}{\bar{Y}^2 c^2 + \sum_{h=1}^{L-1} N_h S_{yh}^2 / N^2}. \quad (1)$$

In this expression one can write $n = n(b_1, \dots, b_L)$ to highlight that the value of n depends on the stratum boundaries. The `strata.LH` function tries to find the optimal boundaries b_h that minimize $n(b_1, \dots, b_{L-1})$. Two minimization algorithms are available, either Sethi's (1963) algorithm as implemented by Lavallée and Hidiroglou (1988) with `algo="Sethi"` or Kozak's (2004) random search algorithm with `algo="Kozak"`. The latter is the default option. This section assumes $Y = X$; it does not distinguish the stratification from the survey variable.

3.1 Sethi (1963) example with the normal distribution

A classical problem is to determine the optimal boundaries for L strata in an infinite population from a known distribution. For instance, Sethi (1963) derived the optimal bounds for the normal and the χ_{30}^2 distributions. To obtain approximate solutions, one can run the `strata.LH` function on a large Monte Carlo population simulated from the known distribution, without requesting a take-all stratum. In (1), one has $N_h / N^2 \approx 0$ and the optimal boundaries are the same for any target CV c .

The following R-code simulates populations of size 10^5 from the $N(10, 1)$ and the χ_{30}^2 distributions. Observe that stratification requires the stratification variable to be non negative, so that it would not work on standard normal deviates. By subtracting 10 from the $N(10, 1)$ boundaries, we get the ones for a $N(0, 1)$. The calculations are done with the `strata.LH` function with the argument `algo="Sethi"` and with `takeall=0`, so that a take-all stratum is not requested.

```
> z <- rnorm(100000, 10)
> z15 <- strata.LH(x = z, CV = 0.001, Ls = 5,
+   alloc = c(0.5, 0, 0.5), takeall = 0, algo = "Sethi")
> z15$bh - 10

[1] -1.1247340 -0.3480829 0.3297044 1.0979017

> x30 <- rchisq(100000, 30)
> x15 <- strata.LH(x = x30, CV = 0.01, Ls = 5,
+   alloc = c(0.5, 0, 0.5), takeall = 0, algo = "Sethi")
> x15$bh

[1] 22.82148 28.12303 33.38642 40.20165
```

In Table 5, the agreement between the true bounds reported in Table 8 of Sethi (1963) and the Monte Carlo bounds is quite good. This approach could be used to calculate the optimal stratum boundaries for an arbitrary distribution, see for instance Khan, Nand, and Ahmad (2008).

Table 5
Comparison of Sethi's (1963) optimal stratum boundaries and of the approximate boundaries obtained with stratification

		stratification's results				Sethi's results			
		1	2	3	4	1	2	3	4
$N(0,1)$	2	-0.007				0.00			
	b_h 3	-0.531	0.567			-0.55	0.55		
	4	-0.883	-0.008	0.864		-0.88	0.00	0.88	
	5	-1.125	-0.348	0.330	1.098	-1.11	-0.34	0.34	1.11
b_h χ^2_{30}	2	30.674				30.6			
	3	26.535	35.141			26.0	35.0		
	4	24.340	30.733	38.179		24.0	30.6	38.0	
	5	22.821	28.123	33.386	40.202	22.0	28.0	33.0	40.0

3.2 Gunning and Horgan (2004) example

In their original proposal, Lavallée and Hidiroglou (1988) always had a take-all stratum for a skewed survey

variable. To show that this was not always mandatory, Gunning and Horgan (2004) derived the optimal stratified designs featuring a take-all stratum for the four populations considered in Table 1. The findings of their Table 7 (with slight corrections due to rounding errors) is reproduced in Table 6. Comparing Tables 1 and 6, one sees that the optimal designs featuring a take-all stratum have n -values larger than 100 for three populations out of four. The optimal design is superior to the geometric design only for the Debtors population. The R-code to run Sethi's algorithm on the Debtors population is given below.

```
pop1LH <- strata.LH(x = Debtors, CV = 0.0359, Ls = 5,
  alloc = c(0.5, 0, 0.5), takeall = 1, algo = "Sethi")
```

In Table 6, one would expect the optimal designs obtained through an iterative algorithm to have a smaller sample size than the *ad hoc* geometric designs. This fails to occur for three populations. This might be caused by a failure of Sethi's algorithm to find the true minimum value for n . To check this, we reran the programs to produce Table 6 with the argument `algo="Kozak"`. The sample sizes n are given in the second column of Table 7. Kozak's algorithm finds a smaller n -value than Sethi's for three of the four populations. This highlights the weakness of Sethi's algorithm for real populations. The second column of Table 7 has n values larger than 100 for two of the four populations. In these cases, the geometric design might be better because a take-all stratum is not required. To check this we reran Kozak's algorithm without a take-all stratum, *i.e.*, with `takeall=0`. The results are reported in the third column of Table 7. For the Debtors and the UScolleges populations, taking away the take-all stratum reduces the sample size n . Still, for the UScities population, Kozak's algorithm does worse than the geometric design. It failed to find the true minimum value of n with the default arguments that control its random search. To better understand the results of Table 7, we now present in more details the selection of initial stratum boundaries in `strata.LH` and the parameters that control the random search with `algo="Kozak"`.

Table 6
Optimal stratified designs featuring a take-all stratum obtained with Sethi's algorithm for the 4 populations of Table 1

Population	n	CV		1	2	3	4	5
Debtors	93	0.0359	b_h	349.33	1,190.16	3,482.98	10,322.50	
			N_h	1,856	991	350	146	26
			n_h	13	17	17	20	26
UScities	137	0.0145	b_h	14.72	21.62	35.59	80.47	
			N_h	189	270	336	164	79
			n_h	4	8	16	30	79
UScolleges	107	0.0183	b_h	512.32	869.76	1,577.23	3,668.85	
			N_h	133	180	185	110	69
			n_h	4	6	10	18	69
USbanks	104	0.0107	b_h	99.37	129.60	181.94	317.36	
			N_h	70	66	82	65	74
			n_h	4	4	7	15	74

Table 7
Sample size n for three optimal designs and four populations

Population	algo=Sethi takeall=1	algo=Kozak takeall=1	algo=Kozak takeall=0
Debtors	93	92	82
UScities	137	114	123
UScolleges	107	107	95
USbanks	104	88	88

3.3 Customization of the algorithms

The default initial stratum boundaries for the two iterative algorithms are the arithmetic starting point of Gunning and Horgan (2007), with $b_h = \min X + (\max X - \min X) \times h/L$, for $h = 1, \dots, L-1$. In Table 7, this choice is questionable and the geometric stratum boundaries would have been closer to the true optimal boundaries. In `strata.LH`, the argument `initbh=` allows to specify a vector of $L-1$ initial boundary values. The maximum number of iterations can be changed with the `maxiter` element of the `algo.control` argument.

Kozak's algorithm was first proposed in Kozak (2004), see also Kozak and Verma (2006). It uses a random search that selects the $L-1$ stratum boundaries among the sorted values of X , with the duplicates discarded. At one iteration, it randomly picks a number d in the set $\{-\text{maxstep}, -\text{maxstep}+1, \dots, \text{maxstep}\}$ and one of the $L-1$ boundaries. Then it moves the selected boundary by d positions in the vector of sorted X -values. If (1) is smaller with the new boundary it is kept, otherwise it is discarded and the boundaries are left unchanged at this iteration. The algorithm stops when the boundaries have not been changed for `maxstill` consecutive iterations. The default values are `maxstep=3` and `maxstill=100`. Two consecutive runs of Kozak's algorithm might lead to different designs because of the random nature of this algorithm. The `strata.LH` runs the algorithm `rep` times and the information for each run is contained in the `rep.detail` element of R-objects of class `strata`; the default value is `rep=3`. If the `rep` runs lead to different designs, then the tuning parameters of the algorithm can be changed. One can also use `rep="change"` which runs the algorithm 27 times with different starting and `maxstep` values. An additional example illustrating an instance where Kozak's algorithm does not reach a global minimum is presented in the Appendix.

With N_u unique X -values, there are approximately $\binom{N_u-1}{L-1}$ possible sets of stratum boundaries. If this number is smaller than `minsol` all the possible sets of strata are tried, rather than carrying out a random search. The default value is `minsol=1000`. The elements `maxstep`, `maxstill`, `minsol` and `rep` belong to the `algo.control` argument. In Table 7, we were unable to improve the geometric stratified design for the UScities population. The command to run

Kozak's algorithm 27 times with various tuning parameters is given below.

```
> data(UScities)
> pop2LHrep <- strata.LH(x = UScities, CV = 0.0145, Ls = 5,
  alloc = c(0.5, 0, 0.5), takeall = 0, algo = "Kozak",
  algo.control = list(rep = "change"))
```

This command takes a few seconds to run and yields a stratified design with $n = 100$, similar to that presented in Table 1 for the UScities.

3.4 Designs with a predetermined sample size n

With Kozak's algorithm it is possible to find the boundaries that minimize the CV of \bar{y}_s for a fixed sample size n rather than minimizing n for a predetermined CV. As an example we revisit the stratified designs of Table 1. The geometric boundaries are used as initial values and the default Kozak algorithm is run. The R-code for the Debtors population is given below.

```
> pop1k <- strata.LH(x = Debtors, initbh = pop1$bh, n = 100,
  Ls = 5, alloc = c(0.5, 0, 0.5), algo = "Kozak")
```

The CVs of the estimator of \bar{y}_s obtained with the optimal stratified designs are 3.12%, 1.43%, 1.72%, and 1.04% for the four populations as compared with 3.59%, 1.45%, 1.83%, and 1.07% in Table 1. Thus the iterative algorithm allowed to reduce the CVs.

4. Stratification with anticipated moments

A difference between the stratification variable X and the survey variable Y can be accounted for by having a model for the conditional distribution of Y given X . In stratification, there is a log-linear model where

$$Y = \exp(\alpha)X^\beta \exp(\sigma\epsilon),$$

and an heteroscedastic linear model with

$$Y = \alpha + \beta X + \sigma\epsilon X^\gamma, \quad (2)$$

and α , β , and γ are real parameters specified by the user and ϵ is a $N(0, 1)$ random variable. A random replacement model (Rivest 1999) is also available and stratum specific mortality rates (Baillargeon, Rivest and Ferland 2007) can be added to the log-linear model.

Under these models, the anticipated mean of Y for the units classified in stratum h , with $X \in [b_{h-1}, b_h)$ are

$$\bar{Y}_h = \frac{1}{N_h} \sum_{b_{h-1} \leq X_i < b_h} E(Y | X_i)$$

while the anticipated variance is

$$S_{yh}^2 = \frac{1}{N_h} \sum_{b_{h-1} \leq X_i < b_h} \{E(Y | X_i) - \bar{E}(Y | X)_h\}^2 + \frac{1}{N_h} \sum_{b_{h-1} \leq X_i < b_h} \text{Var}(Y | X_i)$$

where $\bar{E}(Y | X)_h$ is the average of the predicted values of Y for the units in stratum h . In `strat.cumrootf`, `strata.geo` and `strata.bh` these expressions are used to evaluate the sampling properties of \bar{y}_s while in `strata.LH`, the minimization of (1) is carried out with anticipated moments. In `strata.LH` the stratum boundaries depend on the model for the relationship between X and Y ; they do not for the other `strata` functions.

4.1 An example with the MU284 Swedish municipalities

In Section 2.5 two stratified sampling plans were derived for the MU284 population with *REV84* as stratification variable. The R-code that follows investigates the performance of these sampling designs for the variable *RMT85*. The vector `ord` contains the position of the order statistics of the *REV84* variable; thus `Y[ord]` is the vector of the *RMT85* variable, ordered by increasing *REV84*-value.

```
> data(Sweden)
> X <- Sweden$REV84
> Y <- Sweden$RMT85
> ord <- order(X)
> geo_rmt <- var.strata(geo_cer, y = Y[ord])
> cum_rmt <- var.strata(cum_cer, y = Y[ord])
> c(geo_rmt$RRMSE, cum_rmt$RRMSE)
```

```
[1] 0.06889558 0.07368794
```

In section 2.4, the CVs of the estimator \bar{y}_s for the stratification variable *REV84* were less than 5% for the $\text{cum}\sqrt{f}$ and the geometric designs. When estimating the mean of *RMT85*, the CVs are larger than 6%. This emphasizes that calculating sample sizes with a stratification variable underestimate the n needed to reach the target CV for a different survey variable. These results are reported in the first two designs of Table 8. Table 8 also shows the optimal design calculated by applying Kozak's algorithm to the *REV84* variable, assuming $Y = X$.

Following Rivest (2002), a log-linear model is fitted for the relationship between the two variables. As shown in Figure 2, there are outliers and the following R-code estimates the parameters of the log-linear model by discarding the municipalities with extreme X/Y quantiles. The 18 discarded municipalities are represented by a star in Figure 2. The R-code for fitting the model to the non outliers follows.

```
> keep <- (X/Y > quantile(X/Y, 0.03)) & (X/Y < quantile(X/Y, 0.97))
> reg <- lm(log(Y)[keep] ~ log(X)[keep])
> coef(reg)
```

```
(Intercept) log(X)[keep]
-3.153025 1.058355
```

```
> summary(reg)$sigma
```

```
[1] 0.25677
```

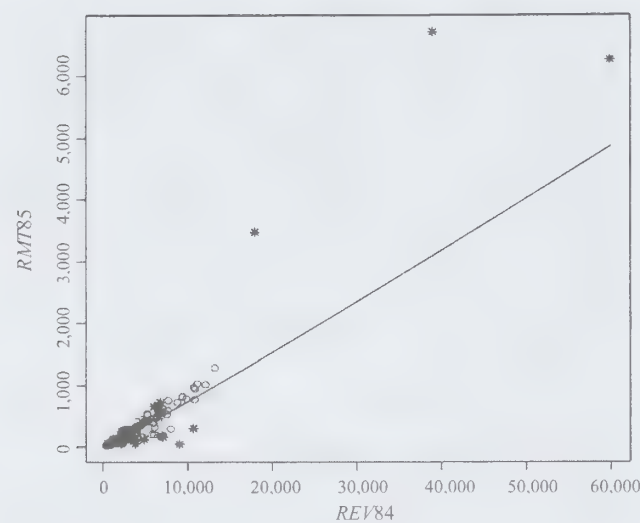



Figure 2 Plot of *RMT85* by *REV84* from the data set Sweden

The following code stratifies the *MU284* population on *REV84* using the $\text{cum}\sqrt{f}$ and the geometric method. The allocation is however carried out with anticipated moments calculated with the log-linear regression model of *RMT85* on *REV84*. The strata of these two designs are the same as those calculated earlier. The model affects only the anticipated CV. It is not so for the optimal design where the anticipated moments are used in the stratification algorithm. Kozak’s algorithm might fail to find the global minimum n value when using anticipated moments; thus we use the bounds calculated with $Y = X$ as starting values.

```
> geo_cer.m <- strata.geo(x = X[ord], CV = 0.05, Ls = 4,
  alloc = c(0.35, 0.35, 0), model = "loglinear",
  certain = (length(X) - 2):length(X), model.control =
  list(beta = 1.058355, sig2 = 0.25677^2))
> geo_cer.var <- var.strata(geo_cer.m, y = Y[ord])
> cum_cer.m <- strata.cumrootf(x = X[ord], nclass = 50,
  CV = 0.05, Ls = 4, alloc = c(0.35, 0.35, 0),
  certain = (length(X) - 2):length(X), model = "loglinear",
  model.control = list(beta = 1.058355, sig2 = 0.25677^2))
> cum_cer.var <- var.strata(cum_cer.m, y = Y[ord])
> LH <- strata.LH(x = X, CV = 0.05, Ls = 5,
  alloc = c(0.35, 0.35, 0), takeall = 1)
> LH.var <- var.strata(LH, y = Y)
> LH_m <- strata.LH(x = X, CV = 0.05, Ls = 5,
  initbh = LH$bh, alloc = c(0.35, 0.35, 0), takeall = 1,
  model = "loglinear", model.control = list(beta = 1.058355,
  sig2 = 0.25677^2))
> LH_m.var <- var.strata(LH_m, y = Y)
```

In Table 8, sample sizes calculated with anticipated moments give CVs smaller than 5% for estimating the mean *RMT85* variable. The optimal LH design requires a n slightly smaller than the other two. Accounting for $Y \neq X$ when minimizing (1) gives a larger take-all stratum since its size increased from 4 to 5 when using the anticipated moments.

Finally observe that the arguments *model* and *model.control* can be used with *var.strata*. For the geometric design considered in this section, one can get results very similar to those obtained with the argument

$y=Y$. As shown below, the model yields a CV of 6.894% as compared with 6.890% obtained with the original *RMT85*-variable. For the $\text{cum}\sqrt{f}$ method the model CV is 7.282% as compared to 7.369% found earlier while for the Lavallée Hidiroglou algorithm these two values are 7.080% and 7.110%.

```
> geo_rmt2 <- var.strata(geo_cer, model = "loglinear",
  model.control = list(beta = 1.058355, sig2 = 0.25677^2))
> geo_rmt2$RRMSE

[1] 0.0689368
```

Table 8
Three stratified designs for estimating the mean *RMT85* with *REV84* as the stratification variable

Model	Method		1	2	3	4	5	n	anticip. CV
$Y = X$	$\text{cum}\sqrt{f}$	N_h	127	79	46	29	3		
		n_h	3	4	4	5	3	19	7.37
	geometric	N_h	42	116	88	35	3		
		n_h	2	5	7	7	3	24	6.89
	LH	N_h	120	82	45	33	4		
		n_h	3	4	4	5	4	20	7.11
loglinear	$\text{cum}\sqrt{f}$	N_h	127	79	46	29	3		
		n_h	6	8	9	10	3	36	4.78
	geometric	N_h	42	116	88	35	3		
		n_h	3	8	13	13	3	40	4.74
	LH	N_h	121	81	45	32	5		
		n_h	6	7	7	9	5	34	4.90

4.2 Anderson, Kish and Cornell (1976) example with the bivariate normal distribution

Anderson *et al.* (1976) investigated the optimal stratification for Y based on X when (X, Y) has a bivariate normal distribution with correlation ρ . Thus model (2) holds with $\alpha = \gamma = 0$, $\beta = \rho$, and $\sigma^2 = 1 - \rho^2$ where X has a $N(0, 1)$ distribution. To reproduce Anderson *et al.* (1976) results, we generate a population of size $N = 10^5$ from a $N(0, 1)$ distribution and select *model*="linear" (as in Section 3.1 a mean of 10 was used to prevent X from being negative). For a linear model, only Kozak’s algorithm works. Given the special nature of the problem, the *maxstep* parameter is set to 20 and only one repetition (*rep*=1) of the algorithm is run. When there is no take-all stratum, the optimal stratum boundaries are independent of the CV, as in Section 3.1. We used $CV = 0.01$ in the calculations.

```
> x <- rnorm(1e+05, 10)
> bi3a <- strata.LH(x = x, CV = 0.01, Ls = 3, takenone = 0,
  model = "linear",
  model.control = list(beta = 0.25, sig2 = 1 - 0.25^2,
  gamma = 0), algo.control = list(maxstep = 20, rep = 1))
> bi3a$bh - 10

[1] -0.619354 0.604198
```

In Table 9, *stratification*'s results are equal to Anderson's *et al.* (1976) findings up to nearly two decimals. This highlights the flexible nature of the package; it can find the optimal stratified design for any distribution of the stratification variable and for some general models for the conditional distribution of Y given X .

Table 9
Comparison of Anderson *et al.* (1976) optimal stratum boundaries with the approximate boundaries obtained with *stratification*

L	$ \rho $	stratification's results				Anderson <i>et al.</i> 's results			
		1	2	3	4	1	2	3	4
3	0.250	-0.619	0.604			-0.61	0.61		
	0.950	-0.591	0.568			-0.58	0.58		
	0.990	-0.571	0.549			-0.56	0.56		
4	0.250	-0.984	0.004	0.985		-0.98	0.00	0.98	
	0.950	-0.930	0.009	0.942		-0.93	0.00	0.93	
	0.990	-0.902	-0.001	0.895		-0.90	0.00	0.90	
5	0.250	-1.245	-0.377	0.387	1.251	-1.24	-0.38	0.38	1.24
	0.950	-1.187	-0.358	0.372	1.197	-1.19	-0.37	0.37	1.19
	0.990	-1.136	-0.344	0.353	1.144	-1.14	-0.35	0.35	1.14

5. Additional features

Baillargeon and Rivest (2009) considered additional aspects of a stratified design, namely stratum specific anticipated non-response rates and the addition of a take-none stratum with a null sample size. This section discusses briefly how these additional items are handled in *stratification*. Non-response needs to be accounted for when optimizing for n . A take-none stratum makes \bar{y}_s biased; in this case the precision target is specified in terms of a Relative Root Mean Squared Error (RRMSE) rather than a CV. Formula (4.3) of Baillargeon and Rivest (2009) provides a generalization of (1) that includes these two features. This is the formula used for calculating sample sizes in the optimization procedure.

5.1 Non-response

Non-response can be corrected *a posteriori*, by dividing the no non-response stratum sample sizes by the response rates. This is illustrated in the following R-code that considers the MRTS variable, representative of Statistics Canada Monthly Retail Trade Survey. *Post hoc* non-response corrections are implemented in the `var.strata` function with the argument `rh.postcorr=TRUE`. An alternative is to consider response rates when allocating the sample to the strata. They can be specified in a `strata` function with the argument `rh=`. This approach penalizes strata with a high non-response; it typically yields a smaller

n value than the *a posteriori* corrections. This is illustrated in the `cum \sqrt{f}` portion of Table 10. With four strata and response rates of 0.8, 0.8, 0.9, 1, the *a posteriori* correction needs $n = 445$ to reach the target CV for the MRTS variable, as compared with $n = 444$ for an allocation that takes non-response into account.

```
> data(MRTS)
> cum <- strata.cumrootf(x = MRTS, nclass = 500, CV = 0.01,
  Ls = 4, alloc = c(0.5, 0, 0.5))
> cum.var <- var.strata(cum, rh = c(0.8, 0.8, 0.9, 1))
> cum.post <- var.strata(cum, rh = c(0.8, 0.8, 0.9, 1),
  rh.postcorr = TRUE)
> cum_rh <- strata.cumrootf(x = MRTS, nclass = 500, CV = 0.01,
  Ls = 4, alloc = c(0.5, 0, 0.5), rh = c(0.8, 0.8, 0.9, 1))
```

Non-response can also be accounted for when constructing an optimal sampling design, either *a posteriori* or in the stratum construction. These two approaches are implemented for the MRTS population in the following R-code. The higher non-response rates for the small units penalize the first stratum which is smaller when non-response is accounted for in the stratification algorithm, as can be seen in Table 10. Still accounting for non-response in the stratum construction gives a smaller n -value than an *a posteriori* correction. Table 3 of Baillargeon and Rivest (2009) presents additional examples, including both anticipated moments and non-response, of the construction of stratified designs for the MRTS population.

```
> LH <- strata.LH(x = MRTS, CV = 0.01, Ls = 4,
  alloc = c(0.5, 0, 0.5), takeall = 1)
> LH.var <- var.strata(LH, rh = c(0.8, 0.8, 0.9, 1))
> LH.post <- var.strata(LH, rh = c(0.8, 0.8, 0.9, 1),
  rh.postcorr = TRUE)
> LH_rh <- strata.LH(x = MRTS, CV = 0.01, Ls = 4,
  alloc = c(0.5, 0, 0.5), takeall = 1, rh = c(0.8, 0.8, 0.9, 1))
```

Table 10
Two examples of non-response correction: Either *a posteriori* (post) or when constructing the design

Method	rh		1	2	3	4	n	anticip. CV
cum \sqrt{f}	none	N_h	778	742	355	125		
		n_h	87	90	88	125	390	1.11
		n_h^{post}	109	113	98	125	445	1.00
	given	N_h	778	742	355	125		
		n_h	105	108	106	125	444	1.00
	LH							
LH	none	N_h	774	675	374	177		
		n_h	77	65	60	177	379	1.11
		n_h^{post}	96	81	67	177	421	1.00
	given	N_h	675	677	449	199		
		n_h	70	69	80	199	418	1.00

5.2 Take-none stratum

A take-none stratum with a null sample size might be advantageous when the population has small units with Y -values close to 0. The precision of \bar{y}_s is then measured by the mean squared error, $\text{Var}(\bar{y}_s) + (T_{0y}/N)^2$, where T_{0y} is

the anticipated Y -total in the take-none stratum. Setting `takenone=1` in the `strata.LH` function constructs an optimal design with a take-none stratum. Baillargeon and Rivest (2009) showed that Sethi's algorithm does not work in this case and that Kozak's algorithm should be used. When a take-none stratum is used, a rough bias correction can be implemented by dividing \bar{y}_s by the proportion of the total of the X variable in the take some strata. Thus the bias penalty in the mean square error might be too stringent and an alternative measure of precision, such as $\text{Var}(\bar{y}_s) + (p \times T_{0v}/N)^2$, could be used in the stratification algorithm where p is a number in $(0, 1)$. This smaller bias penalty can be implemented by setting the argument `bias.penalty` equal to p . The following R-code constructs three optimal stratified designs for the MRTS population, with and without a take-none stratum; the default full bias penalty is compared to a reduced penalty with $p = 0.5$.

```
> data(MRTS)
> notn <- strata.LH(x = MRTS, CV = 0.1, Ls = 3,
  alloc = c(0.5, 0, 0.5))
> tnl <- strata.LH(x = MRTS, CV = 0.1, Ls = 3,
  alloc = c(0.5, 0, 0.5), takenone = 1)
> tn0.5 <- strata.LH(x = MRTS, CV = 0.1, Ls = 3,
  alloc = c(0.5, 0, 0.5), takenone = 1, bias.penalty = 0.5)
```

The sample sizes n for the three designs are given in Table 11. Including a take-none stratum with a full bias penalty reduces n , from 22 to 16; for this design the take-none stratum accounts for 3% of the total of the X -variable. Reducing the bias penalty to $p = 0.5$ increases the size of the take-none stratum and reduces n . Additional illustrations are given in Table 2 of Baillargeon and Rivest (2009). They show that the size of a take-none stratum typically decreases with the target RRMSE. For the MRTS example, the addition of a take-none stratum diminishes the n -value substantially while for others it does not change the design.

Table 11
Sample sizes for three optimal stratified designs for the MRTS population

takenone	0	1	1
bias.penalty	NA	1	0.5
n	22	16	13
$\% T_x$	0	3	9

6. Conclusion

The R-package *stratification* offers flexible methods for the construction of a stratified sampling design using a univariate stratification variable such as a measure of size in a business survey. Several methods are available to determine the stratum boundaries and the stratum sample sizes.

stratification allows the investigation of features such as a take-all stratum, a take-none stratum, the extent of the discrepancy between X and Y , and a stratum specific non-response.

Acknowledgements

We are grateful to S. Er, E. Gagnon, M. Kozak, and J. Stardom for constructive comments on the package and to the Canada Research Chair on Statistical Sampling and Data Analysis and the Natural Sciences and Engineering Research Council of Canada for their financial support. This research was supported by U.S. National Science Foundation grant SES-0751671.

7. Appendix

7.1 More details on Kozak's algorithm

As described in Section 3.3 Kozak's algorithm uses a random search. Besides decreasing the optimization criterion, either the n -value or the RRMSE of \bar{y}_s , *stratification* requires that the take-some strata contain at least `minNh` units and that they have positive sample sizes, for the new boundary to be admissible. The default is `minNh=2`. A non random, Kozak's algorithm is also available with `method="modified"` in the `algo.control` argument. It tries all the possible changes at one iteration and picks the one that gives the largest drop of the optimization criterion. It is slower than Kozak's algorithm without improving the detection of the global minimum of the optimization criterion. Therefore, it will not be discussed any further.

To illustrate the complete enumeration of all possible solutions mentioned in Section 3.3, consider the `USbanks` data set. It contains 357 values, but only 200 unique values. If one wishes to stratify this population in two strata, only $\binom{200-1}{2-1} = 199$ solutions are possible. The following command performs a complete enumeration of the possible solutions:

```
> enum <- strata.LH(x = USbanks, CV = 0.05, Ls = 2,
  alloc = c(0.5, 0, 0.5))
```

These solutions, with their associated optimization criteria value, can be found in `enum$sol.detail`. Only the solutions fulfilling the admissibility constraints mentioned above are included in `enum$sol.detail`.

When running Kozak's algorithm, the initial boundary values might fail to meet the admissibility constraints; the algorithm might not be able to move at all. In such a case, the initial boundaries are replaced by robust ones. The robust boundaries give an empty take-none stratum if such a stratum is requested, take-all strata as small as possible, and take-some strata with approximately the same number of unique X -values.

Consider once again the example of Section 3.2 with the `UScities` data set, where Kozak's algorithm reached a local minimum with the default arguments. With geometric initial boundaries, Kozak's algorithm converges rapidly to what appears to be a global minimum.

```
> LH_init <- strata.LH(x = UScities, initbh = pop2$bh,
  n = 100, Ls = 5, alloc = c(0.5, 0, 0.5), takeall = 0,
  algo.control = list(rep = 1))
> LH_init$iter.detail
```

	b1	b2	b3	b4	opti	step	iter	run
1	18.5	33.5	59.5	107	0.01444981	0	0	1
2	20.5	33.5	59.5	107	0.01435576	2	2	1
3	19.5	33.5	59.5	107	0.01434272	-1	10	1
4	19.5	33.5	58.0	107	0.01432714	-1	12	1
5	19.5	31.5	58.0	107	0.01431013	-2	13	1
6	19.5	32.5	58.0	107	0.01430163	1	63	1

```
> LH_init$niter
[1] 163
```

The output element `LH_init$iter.detail` contains information about the initial boundaries and the 5 iterations with a change of boundaries only. A total of 163 iterations were needed for the algorithm to converge. The geometric initial boundaries are very close to the optimal solutions. A local minimum can also be avoided by changing some of the algorithm's parameters. The following R-code allows larger steps (`maxstep=20`) and increases the maximal number of iterations (`maxstill=1000`) and the number of repetitions of the algorithm (`rep=20`).

```
> LH_param <- strata.LH(x = UScities, n = 100, Ls = 5,
  alloc = c(0.5, 0, 0.5), takeall = 0, algo.control =
  list(maxstep = 20, maxstill = 1000, rep = 20))
```

The results for the 20 repetitions are reported in `LH_param$rep.detail` and summarized in Table 12. The

solution obtained with the geometric initial boundaries is reached 9 times out of 20.

Table 12
Solutions found by Kozak's algorithm for 20 repetitions

CV	B1	B2	B3	B4	frequency
0.0143	19.50	32.50	58.00	107.00	9
0.0167	16.50	23.50	37.50	78.00	5
0.0167	15.50	22.50	35.50	73.00	6

Figure 3 shows how larger steps help the algorithm to reach the global minimum ($CV=0.0143$), compared to a run of the algorithm with the default arguments (dotted lines, $CV=0.0167$).

7.2 R package stratification summary table

This appendix provides a quick reference for the R package *stratification*. Table 13 lists the five functions in *stratification* and their arguments. The following notes complete the table.

(1) According to the general allocation scheme (Hidiroglou and Srinath 1993). The stratum sample sizes are proportional to $N_h^{2q_1} \bar{Y}_h^{2q_2} S_{yh}^{2q_3}$.

(2) The default value of `initbh` is the set of arithmetic starting points of Gunning and Horgan (2007), see Section 3.3. If `takenone=1` and `initbh` is of size `Ls-1`, the initial boundary of the take-none stratum is set to the first percentile of X . If this first percentile is equal to the minimum value of X , this initial boundary would lead to an empty take-none stratum. In that case, the initial boundary of the take-none stratum is rather set to the second smallest value of X .

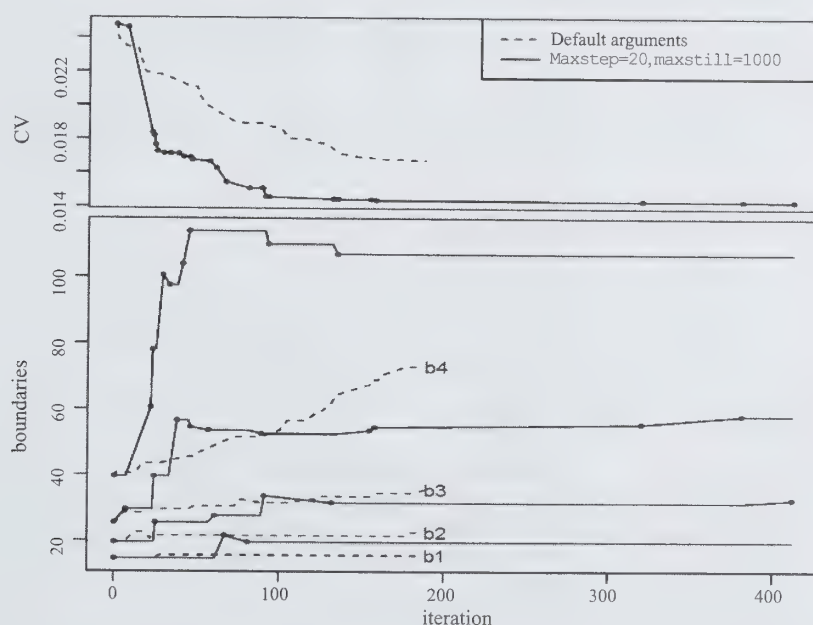


Figure 3 Iterations histories for two runs of Kozak's algorithm

(3) The elements to specify in the `algo.control` argument depend on the algorithm. The following table shows the elements used by each algorithm and their default values. See `help(strata.LH)` for a complete description of every element.

Algorithm	maxiter	method	minNh	maxstep	maxstill	rep	minsol
Sethi	500	-	-	-	-	-	-
Original Kozak	10,000	"original"	2	3	100	3	1,000
Modified Kozak	3,000	"modified"	2	3	-	-	1,000

(4) The elements of the `model.control` argument depend on the model:

- loglinear model with mortality:

$$Y = \begin{cases} \exp(\alpha + \beta \log(X) + \epsilon) & \text{with probability } p_h \\ 0 & \text{with probability } 1 - p_h \end{cases}$$

where $\epsilon \sim N(0, \text{sig}2)$ is independent of X . The parameter p_h is specified through `ph`, `ptakenone` and `pcertain`.

- heteroscedastic linear model :

$$Y = \beta X + \epsilon$$

where

$$\epsilon \sim N(0, \text{sig}2 X^{\text{gamma}}).$$

- random replacement model:

$$Y = \begin{cases} X & \text{with probability } 1 - \epsilon \\ X_{\text{new}} & \text{with probability } \epsilon \end{cases}$$

where X_{new} is a random variable independent of X with the same distribution as X .

The following table presents `model.control` default values according to the model.

model	beta	sig2	ph	ptakenone	pcertain	gamma	epsilon
"loglinear"	1	0	<code>rep(1, Ls)</code>	1	1	-	-
"linear"	1	0	-	-	-	0	-
"random"	-	-	-	-	-	-	0

Table 13
R package *stratification* summary table

argument	Strata.cumrootf	Strata.geo	Strata.LH	Strata.bh	Var.strata	description	format	default
<code>x</code>	•	•	•	•		stratification variable	vector	none (<code>x</code> is mandatory)
<code>n</code>	•	•	•	•		target total sample size	scalar	none (<code>n</code> or <code>CV</code> is mandatory)
<code>CV</code>	•	•	•	•		target CV or RRMSE	scalar	none (<code>n</code> or <code>CV</code> is mandatory)
<code>Ls</code>	•	•	•	•		number of sampled strata	scalar	3
<code>alloc</code>	•	•	•	•		allocation specification (1)	list (<code>q1</code> , <code>q2</code> , <code>q3</code>) where $q_i \geq 0$	Neyman ($q_1 = q_3 = 0.5$, $q_2 = 0$)
<code>certain</code>	•	•	•	•		x -indices for units sampled with certainty	vector	NULL (no certainty stratum)
<code>nclass</code>	•					number of bins	scalar	$\min(10L, N)$
<code>bh</code>				•		strata boundaries	vector	none (<code>bh</code> is mandatory)
<code>takeall.adjust</code>				•		indicator of adjustment for take-all strata	True or False	FALSE (no adjustment)
<code>takeall</code>			•	•		number of take-all strata	one of {0, 1, ..., $L_s - 1$ }	0
<code>initbh</code>			•			initial strata boundaries (2)	vector	equidistant boundaries
<code>algo</code>			•			algorithm identification	"Kozak" or "Sethi"	"Kozak"
<code>algo.control</code>			•			algorithm's parameters specification (3)	list (<code>maxiter</code> , <code>method</code> , <code>minNh</code> , <code>maxstep</code> , <code>maxstill</code> , <code>rep</code>)	depends on algo
<code>strata</code>					•	stratification scheme	strata object	none (<code>strata</code> is mandatory)
<code>y</code>					•	study variable	vector	NULL (model given instead)
<code>model</code>	•	•	•	•	•	model identification	"none", "loglinear", "linear"* or "random"* →	"none"
<code>model.control</code>	•	•	•	•	•	model's parameter specification (4)	list (<code>beta</code> , <code>sig2</code> , <code>ph</code> , <code>ptakenone</code> , <code>gamma</code> , <code>epsilon</code>)	(*unavailable with Sethi's algo) depends on model, but equivalent to <code>model="none"</code>
<code>rh</code>	•	•	•	•	•	anticipated response rates	scalar or vector	<code>rep(1, Ls)</code> or <code>rh</code> from strata
<code>rh.postcorr</code>					•	indicator of posterior correction for non-response	TRUE or FALSE	FALSE (no correction)
<code>takenone</code>			•	•		number of take-none strata	0 or 1	0
<code>bias.penalty</code>			•	•		penalty for the bias	scalar	1

References

- Anderson, D.W., Kish, L. and Cornell, R.G. (1976). Quantifying gains from stratification for optimum and approximately optimum strata using a bivariate normal model. *Journal of the American Statistical Association*, 71, 887-892.
- Baillargeon, S., Rivest, L.-P. and Ferland, M. (2007). Stratification en enquêtes entreprises : une revue et quelques avancées. *Proceedings of the Survey Methods Section, Statistical Society of Canada* (www.ssc.ca/survey/documents/SSC2007_S_Baillargeon.pdf).
- Baillargeon, S., and Rivest, L.-P. (2009). A general algorithm for univariate stratification. *International Statistical Review*, 77, 331-344.
- Cochran, W.G. (1961). Comparison of methods for determining stratum boundaries. *Bulletin of the International Statistical Institute*, 32, 345-358.
- Cochran, W.G. (1977). *Sampling Techniques. Third Edition*. New York: John Wiley & Sons, Inc.
- Dalenius, T., and Hodges, J.L., Jr. (1959). Minimum variance stratification. *Journal of the American Statistical Association*, 54, 88-101.
- Dayal, S. (1985). Allocation of sample using values of auxiliary characteristics. *Journal of Statistical Planning and Inference*, 11, 321-328.
- Detlefsen, R.E., and Veum, C.S. (1991). Design issues for the Retail Trade Sample Surveys of the U.S. Bureau of the Census. *Proceedings of the Survey Research Methods Section, American Statistical Association*, 214-219.
- Gunning, P., and Horgan, J.M. (2004). A new algorithm for the construction of stratum boundaries in skewed populations. *Survey Methodology*, 30, 159-166.
- Gunning, P., and Horgan, J.M. (2007). Improving the Lavallée and Hidiroglou algorithm for stratification of skewed populations. *Journal of Statistical Computation and Simulation*, 77, 277-291.
- Hidiroglou, M.A. (1986). The construction of a self-representing stratum of large units in survey design. *The American Statistician*, 40, 27-31.
- Hidiroglou, M.A., and Srinath, K.P. (1993). Problems associated with designing subannual business surveys. *Journal of Business and Economic Statistics*, 11, 397-405.
- Khan, M.G.M., Nand, N. and Ahmad, N. (2008). Determining the optimum strata boundary points using dynamic programming. *Survey Methodology*, 34, 205-214.
- Kozak, M. (2004). Optimal stratification using random search method in agricultural surveys. *Statistics in Transition*, 6, 797-806.
- Kozak, M., and Verma, M.R. (2006). Geometric versus optimization approach to stratification: A comparison of efficiency. *Survey Methodology*, 32, 157-163.
- Lavallée, P., and Hidiroglou, M. (1988). On the stratification of skewed populations. *Survey Methodology*, 14, 33-43.
- McEvoy, R.H. (1956). Variation in bank asset portfolios. *The Journal of Finance*, 11(4), 463-473.
- Rivest, L.-P. (1999). Stratum jumpers: Can we avoid them?. *ASA Proceedings of the Section on Survey Research Methods, American Statistical Association*, (Alexandria, VA), 64-72.
- Rivest, L.-P. (2002). A generalization of the Lavallée and Hidiroglou algorithm for stratification in business surveys. *Survey Methodology*, 28, 191-198.
- Särndal, C.E., Swensson, B. and Wretman, J. (1992). *Model Assisted Survey Sampling*. New York: Springer Verlag.
- Sethi, V.K. (1963). A note on the optimum stratification of populations for estimating the population means. *Australian Journal of Statistics*, 5, 20-33.
- Sigman, R.S., and Monsour, N.J. (1995). Selecting samples from list frames of businesses. In *Business Survey Methods*, (Eds., B.G. Cox, D.A. Binder, B.N. Chinnappa, A. Christianson, M.L. Colledge and P.S. Kott), 133-152.
- Slanta, J., and Krenzke, T. (1996). Applying the Lavallée and Hidiroglou method to obtain stratification boundaries for the Census Bureau's Annual Capital Expenditure Survey. *Survey Methodology*, 22, 65-75.
- Sweet, E.M., and Sigman R.S. (1995). Evaluation of model-assisted procedures for stratifying skewed populations using auxiliary data. *U.S. Bureau of the Census* (www.census.gov/srd/papers/pdf/sm95-22.pdf).

Replication variance estimation under two-phase sampling

Jae Kwang Kim and Cindy Long Yu¹

Abstract

In two-phase sampling for stratification, the second-phase sample is selected by a stratified sample based on the information observed in the first-phase sample. We develop a replication-based bias adjusted variance estimator that extends the method of Kim, Navarro and Fuller (2006). The proposed method is also applicable when the first-phase sampling rate is not negligible and when second-phase sample selection is unequal probability Poisson sampling within each stratum. The proposed method can be extended to variance estimation for two-phase regression estimators. Results from a limited simulation study are presented.

Key Words: Double sampling; Jackknife; Regression estimator; Reweighted expansion estimator.

1. Introduction

Two-phase sampling, first introduced by Neyman (1938) and sometimes called double sampling, is a cost effective technique in survey sampling. It is typically used when it is very expensive to collect data on the variables of interest, but it is relatively inexpensive to collect data on variables that are correlated with the variables of interest. Two-phase sampling has application in different forms (*e.g.*, Rao 1973; Cochran 1977; Breidt and Fuller 1993; Rao and Sitter 1995; Hidioglou and Särndal 1998; Fuller 1998; Hidioglou 2001; Fuller 2003). Two-phase sampling for stratification refers to the situation where the observation from the first-phase sample is used to make a stratification for the second-phase sampling. By selecting the first-phase sample for stratification purpose, two-phase sampling is a useful tool when there is no sampling frame available for stratification at the beginning. For example, in forest surveys, it is very difficult and expensive to travel to remote areas to make on-ground determinations. However, aerial photographs are relatively inexpensive, and determinations on, say, forest type from aerial photos are strongly correlated with ground determinations and can be used to stratify the first phase sample.

Replication variance estimation is very popular in complex surveys. Rust and Rao (1996) and Wolter (2007) provide comprehensive overviews on this topic. The replication method does not require the computation of the partial derivative of the Taylor expansion and the user can easily produce variance estimates without knowing the sampling design that was used to collect the data. Furthermore, this tendency is increasing because of confidentiality issues (Lu and Sitter 2006). Once the replication weights are provided, the design information such as stratum identifier is not needed for the user's analysis.

There are two commonly used estimators of the population mean under two phase sampling: the double expansion

estimator (DEE) and the reweighted expansion estimator (REE), named by Kott and Stukel (1997). In general the REE is more efficient than the DEE in the situation of two-phase sampling for stratification when the y 's within a stratum are homogeneous. Variance estimation for two-phase sampling is a challenging practical problem, and replication variance estimation is of interest among practitioners. Rao and Shao (1992) proposed a consistent jackknife variance estimator for the REE in the context of hot deck imputation treating the respondents as the second-phase sample. Kott and Stukel (1997) considered the same problem and concluded that the jackknife variance estimator works well for the REE if the first-phase sampling rate is negligible. The sampling rate, or the sampling fraction, $f_1 = nN^{-1}$ is called negligible if f_1 converges to zero under the asymptotic setup described in Section 2. Binder, Babyak, Brodeur, Hidioglou and Jocelyn (2000) studied variance estimation for a similar two-phase sample design using the Taylor linearization method. Kim *et al.* (2006, KNF) provided a rigorous investigation of the replication method and considered replication for other types of estimators. The KNF method has been developed mainly under the situation where the first-phase sampling rate is negligible and the second-phase sampling is a stratified random sampling. If the first-phase sampling rate is not negligible, additional replicates are needed to get consistent variance estimates.

In this paper, we propose a new replication method for variance estimation under two-phase sampling. The proposed method is an extension of the KNF method to cover the situation where the first-phase sampling rate is not necessarily negligible. Unlike the KNF method, the proposed method does not require additional replicates for bias correction in the variance estimation, but does require adjustments in the replication weights. Also, the proposed method is applicable to unequal probability Poisson sampling within

1. Jae Kwang Kim, Department of Statistics, Iowa State University, Ames, Iowa 50011, U.S.A.; Cindy Long Yu, Department of Statistics, Center for Survey Statistics and Methodology, Iowa State University, Ames, Iowa 50011, U.S.A. E-mail: cindyyu@iastate.edu.

second-phase strata, which was not discussed in KNF. Because the proposed method is a replication-based method, it is very easy to implement and can be applied to various types of estimators.

The rest of the paper is organized as follows. In Section 2, the basic setup is introduced, and in Section 3, the proposed method is described. In Section 4, the proposed method is extended to other estimators in two-phase sampling. In Section 5, results from a limited simulation study are presented. Concluding remarks are made in Section 6.

2. Basic setup

For better motivation, in this section we simply assume the situation where the first phase is a simple random sample of size n from a finite population of size N and the second phase sampling is a stratified random sample. In section 3, the setup is extended to include any arbitrary measurable sampling in the first phase and unequal probability Poisson sampling within each stratum in the second phase. Using the information obtained from the first-phase sample, it is stratified into H strata for second-phase sampling. In stratum h , we have n_h first-phase sample elements and let A_{h1} be the set of indices for the first-phase sample elements in stratum h . In the second-phase sampling, a stratified random sample of size r is selected with sample size $r_h (\leq n_h)$ in stratum h , where $r = \sum_{h=1}^H r_h$ and the sampling rate r_h/n_h is fixed for each stratum. To formally discuss the asymptotic theory, we assume a sequence of finite populations, a sequence of first-phase samples, and a sequence of second-phase samples, as described in KNF. In this asymptotic setup, we allow that the second-phase sample size r goes to infinity at the same rate as the first phase sample size n , i.e., $r = O(n)$ and $r^{-1} = O(n^{-1})$, and H is fixed. Thus, in the setup of fixed H , $r_h^{-1} = O(n^{-1})$.

When the study variable y_i is observed in the second phase sample, the population mean of y is estimated by

$$\bar{y}_{tp} = \frac{1}{n} \sum_{h=1}^H \sum_{i \in A_{h2}} \frac{n_h}{r_h} y_i,$$

where A_{h2} is the set of indices for the second-phase sample elements that belong to stratum h . The variance of \bar{y}_{tp} can be written as

$$\text{Var}(\bar{y}_{tp}) = \left(\frac{1}{n} - \frac{1}{N} \right) S^2 + E \left\{ \sum_{h=1}^H \left(\frac{n_h}{n} \right)^2 \left(\frac{1}{r_h} - \frac{1}{n_h} \right) s_{h1}^2 \right\} \quad (1)$$

where $\bar{y}_1 = n^{-1} \sum_{h=1}^H \sum_{i \in A_{h1}} y_i$, $S^2 = (N-1)^{-1} \sum_{i=1}^N (y_i - \bar{y}_1)^2$, $s_{h1}^2 = (n_h - 1)^{-1} \sum_{i \in A_{h1}} (y_i - \bar{y}_{h1})^2$, and $\bar{y}_{h1} = n_h^{-1} \sum_{i \in A_{h1}} y_i$. Using

$$n^{-1} S^2 \doteq E \left\{ n^{-1} \sum_{h=1}^H w_h [(\bar{y}_{h1} - \bar{y}_1)^2 + s_{h1}^2] \right\}$$

where $w_h = n^{-1} n_h$ and \doteq indicates an approximation ignoring the terms of order $o(n^{-1})$, the variance term (1) is approximated by

$$\text{Var}(\bar{y}_{tp}) \doteq E \left\{ n^{-1} (1 - f_1) \sum_{h=1}^H w_h (\bar{y}_{h1} - \bar{y}_1)^2 + \sum_{h=1}^H (r_h^{-1} - n_h^{-1} f_1) w_h^2 s_{h1}^2 \right\}, \quad (2)$$

where $f_1 = nN^{-1}$.

A consistent estimator of the variance of \bar{y}_{tp} can be derived from (2) by replacing \bar{y}_{h1} and s_{h1}^2 by their estimates $\bar{y}_{h2} = r_h^{-1} \sum_{i \in A_{h2}} y_i$ and $s_{h2}^2 = (r_h - 1)^{-1} \sum_{i \in A_{h2}} (y_i - \bar{y}_{h2})^2$, respectively. That is, a consistent variance estimator is

$$\hat{V} = n^{-1} (1 - f_1) \sum_{h=1}^H w_h (\bar{y}_{h2} - \bar{y}_2)^2 + \sum_{h=1}^H (r_h^{-1} - n_h^{-1} f_1) w_h^2 s_{h2}^2, \quad (3)$$

where $\bar{y}_2 = \sum_{h=1}^H w_h \bar{y}_{h2}$. The variance estimator (3) is a linearized variance estimator.

Kott and Stukel (1997) and KNF developed a jackknife variance estimator by successively deleting units from the entire first-phase sample and then adjusting the weights. The full jackknife replicates are

$$\bar{y}_{tp}^{(k)} = \frac{1}{N} \sum_{h=1}^H \hat{N}_{h1}^{(k)} \bar{y}_{h2}^{(k)} \quad (4)$$

where k is the index of the unit deleted in the jackknife replicate,

$$\begin{aligned} \frac{1}{N} \hat{N}_{h1}^{(k)} &= \sum_{i \in A_{h1}} w_i^{(k)} \\ &= \begin{cases} (n-1)(n_h-1) & \text{if } k \in A_{h1} \\ (n-1)n_h & \text{if } k \notin A_{h1} \end{cases} \end{aligned}$$

and

$$\begin{aligned} \bar{y}_{h2}^{(k)} &= \frac{\sum_{i \in A_{h2}} w_i^{(k)} y_i}{\sum_{i \in A_{h2}} w_i^{(k)}} \\ &= \begin{cases} (r_h-1)^{-1} (r_h \bar{y}_{h2} - y_k) & \text{if } k \in A_{h2} \\ \bar{y}_{h2} & \text{if } k \notin A_{h2}. \end{cases} \end{aligned} \quad (5)$$

The full jackknife variance estimator of the form

$$\hat{V}_J = \sum_{k \in A_1} \frac{n-1}{n} (1 - f_1) (\bar{y}_{tp}^{(k)} - \bar{y}_{tp})^2, \quad (6)$$

where $\bar{y}_{tp}^{(k)}$ is defined in (4), is asymptotically equivalent to

$$\begin{aligned} \hat{V}_J &\doteq n^{-1} (1 - f_1) \sum_{h=1}^H w_h (\bar{y}_{h2} - \bar{y}_2)^2 \\ &\quad + (1 - f_1) \sum_{h=1}^H r_h^{-1} w_h^2 s_{h2}^2. \end{aligned} \quad (7)$$

Thus, comparing (7) with (2), the bias of the jackknife variance estimator (6) is

$$\text{Bias}(\hat{V}_J) \doteq -E \left\{ f_1 \sum_{h=1}^H (r_h^{-1} - n_h^{-1}) s_{h2}^2 \right\}.$$

Therefore, if the first-phase sampling rate is negligible in the sense of $f_1 \doteq 0$, the bias is negligible, *i.e.*, the bias = $o(n^{-1})$. Otherwise, the variance estimator underestimates the variance.

To consider a bias-corrected jackknife method, instead of (5), we consider

$$\bar{y}_{h2}^{(k)} = \begin{cases} (r_h - \delta_h)^{-1} (r_h \bar{y}_{h2} - \delta_h y_k) & \text{if } k \in A_{h2} \\ \bar{y}_{h2} & \text{if } k \notin A_{h2}, \end{cases} \quad (8)$$

where δ_h is to be determined. In (5), $\delta_h = 1$ was used. The jackknife variance estimator using (8) instead of (5) is asymptotically equivalent to

$$\begin{aligned} \hat{V}_J &\doteq n^{-1} (1 - f_1) \sum_{h=1}^H w_h (\bar{y}_{h2} - \bar{y}_2)^2 \\ &\quad + (1 - f_1) \sum_{h=1}^H \frac{(r_h - 1) \delta_h^2}{(r_h - \delta_h)^2} w_h^2 s_{h2}^2. \end{aligned}$$

Thus, the asymptotic bias is

$$\begin{aligned} \text{Bias}(\hat{V}_J) &\doteq \\ E \left[\sum_{h=1}^H \left\{ (1 - f_1) \frac{(r_h - 1) \delta_h^2}{(r_h - \delta_h)^2} - \frac{1}{r_h} \left(1 - f_1 \frac{r_h}{n_h} \right) \right\} w_h^2 s_{h2}^2 \right]. \end{aligned}$$

The asymptotic bias is zero if

$$\delta_h = \frac{r_h}{1 + \sqrt{r_h (r_h - 1) / d_h}}$$

where $d_h = \sqrt{(1 - f_1 r_h n_h^{-1}) / (1 - f_1)}$. Hence, with such determined δ_h in equation (8), the resulting jackknife variance estimator is approximately unbiased without assuming $f_1 \doteq 0$.

3. Proposed method

The proposed method in Section 2 is now extended to a more general first-phase sampling design. To do this, we need to assume that the replication variance estimator of the form

$$\hat{V}_1 = \sum_{k=1}^L c_k (\hat{\theta}^{(k)} - \hat{\theta})^2,$$

where $\hat{\theta} = \sum_{i \in A_1} w_i y_i$, and $\hat{\theta}^{(k)} = \sum_{i \in A_1} w_i^{(k)} y_i$, is consistent for the variance of $\hat{\theta}$ under the single (first) stage sampling design. That is,

$$\frac{\hat{V}_1}{\text{Var}(\hat{\theta})} - 1 = o_p(1). \quad (9)$$

Here L is the number of replicates. For most of the measurable designs, which are designs with all positive joint inclusion probabilities, we can construct a replication variance estimator satisfying (9) even when the sample rate $f = n/N$ is large. For example, see Fay (1984) and Flyer (1987). Brick and Morganstein (1996) describes the basic algorithm for WesVar, a commercially available software for replication variance estimation in survey sampling.

In this section, we also consider a more challenging case of stratified unequal probability sampling for the second phase. More specifically, the second phase sampling considered is unequal probability Poisson sampling within the second-phase strata. Fuller (1998) also considered Poisson sampling in the second phase and argued that Poisson sampling in the second phase sampling is a good approximation. An example of this in the context of forest surveys is that, in addition to forest types, the photo-interpretors can also identify tree density and tree height from the aerial photos taken in the first phase, which can be used to construct the second phase selection probabilities within each stratum (forest type).

In this section, we will focus on the REE-type estimator first since it is more efficient than the DEE-type, and extension to the DEE is discussed in Section 4. Let w_i be the first-phase sampling weight and let w_{i2} be the inverse of the conditional probability in the second-phase. That is, $w_{i2} = \pi_{i2}^{-1}$ where $\pi_{i2} = \text{Pr}(i \in A_{h2} | i \in A_{h1})$. The REE-type estimator can be written as

$$\bar{y}_{ip} = \frac{1}{N} \sum_{h=1}^H \hat{N}_{h1} \bar{y}_{h2} \quad (10)$$

where $\hat{N}_{h1} = \sum_{i \in A_{h1}} w_i$ and $\bar{y}_{h2} = (\sum_{i \in A_{h2}} w_i \pi_{i2}^{-1})^{-1} \sum_{i \in A_{h2}} w_i \pi_{i2}^{-1} y_i$. In KNF, π_{i2} is assumed to be constant within the second-phase stratum.

We consider a replication-based approach for variance estimation of the REE-type estimator (10) when π_{i2} is not necessarily constant within the second-phase stratum. We consider the special case when the second-phase sampling design is Poisson sampling. Using the replication method satisfying (9), the KNF-type variance estimator can be applied to estimate the variance of \bar{y}_{ip} in this situation. That is,

$$\hat{V}_{\text{KNF}} = \sum_{k=1}^L c_k (\bar{y}_{ip}^{(k)} - \bar{y}_{ip})^2, \quad (11)$$

where

$$\bar{y}_{ip}^{(k)} = \frac{1}{N} \sum_{h=1}^H \hat{N}_{h1}^{(k)} \bar{y}_{h2}^{(k)} \quad (12)$$

with $\bar{y}_{h2}^{(k)} = (\sum_{i \in A_{h2}} w_i^{(k)} \pi_{i2}^{-1})^{-1} \sum_{i \in A_{h2}} w_i^{(k)} \pi_{i2}^{-1} y_i$ and $\hat{N}_{h1}^{(k)} = \sum_{i \in A_{h1}} w_i^{(k)}$, and c_k is a factor associated with replicate k determined by the replication method. Under Poisson sampling in the second phase, we have the following asymptotic bias:

$$\text{Bias}(\hat{V}_{\text{KNF}}) = -\frac{1}{N^2} \sum_{h=1}^H \sum_{i \in U_h} \pi_{i2}^{-1} (1 - \pi_{i2}) (y_i - \bar{Y}_h)^2, \quad (13)$$

where U_h is the set of indices of population elements in stratum h and $\bar{Y}_h = N_h^{-1} \sum_{i \in U_h} y_i$. A sketched proof of (13) is presented in Appendix A.

An asymptotically unbiased estimator of the bias (13) is

$$\hat{V}_{\text{bias}} = -\frac{1}{N^2} \sum_{h=1}^H \sum_{i \in A_{h2}} w_i \pi_{i2}^{-2} (1 - \pi_{i2}) (y_i - \bar{y}_{h2})^2. \quad (14)$$

The bias is negligible if $n/N \doteq 0$. Thus, we can safely ignore the bias of the KNF-type variance estimator when the first-phase sampling rate is negligible. The bias can be arbitrarily large if the first-phase sampling rate n/N is not negligible. KNF also discuss a bias-correction replication method using additional replicates, which can lead to a large number of replicates. Creating additional replicates for bias-correction can be cumbersome for large scale surveys.

We consider an alternative bias-corrected replication variance estimator that does not require creating additional replicates. To develop a replication-based bias-corrected variance estimator, define a random variable

$$\delta_{ki}^{\text{indep}} \sim \text{Bernoulli}(p_k), \quad (15)$$

where p_k is to be determined. Let

$$\hat{V}_{\text{KNF}}^* = \sum_{k=1}^L c_k (\bar{y}_{ip}^{*(k)} - \bar{y}_{ip})^2 \quad (16)$$

where

$$\bar{y}_{ip}^{*(k)} = \frac{1}{N} \sum_{h=1}^H \hat{N}_{h1}^{(k)} \bar{y}_{h2}^{*(k)} \quad (17)$$

with $\hat{N}_{h1}^{(k)} = \sum_{i \in A_{h1}} w_i^{(k)}$,

$$\bar{y}_{h2}^{*(k)} = \frac{\sum_{i \in A_{h2}} w_i^{(k)} M_{i2}^{(k)} \pi_{i2}^{-1} y_i}{\sum_{i \in A_{h2}} w_i^{(k)} M_{i2}^{(k)} \pi_{i2}^{-1}} \quad (18)$$

with

$$M_{i2}^{(k)} = 1 + (\delta_{ki} - p_k) b_i \quad (19)$$

and b_i is also to be determined. By construction, $E_*(\delta_{ki} - p_k) = 0$, where E_* denotes that the expectation is taken with respect to the mechanism in (15). Thus, the replicates (18) create additional variation in the replication weights, where the additional variation in (18) comes from

the distribution (15). A suitable choice of p_i and b_i can make the resulting variance estimator consistent.

Under the regularity conditions discussed in KNF, we have

$$E_*(\hat{V}_{\text{KNF}}^*) = \hat{V}_{\text{KNF}} + N^{-2} \sum_{h=1}^H \sum_{i \in A_{h2}} w_i^2 b_i^2 \pi_{i2}^{-2} u(y_i - \bar{y}_{h2})^2 + o_p(n^{-1}), \quad (20)$$

where $u = \sum_{k=1}^L c_k p_k (1 - p_k)$. A sketched proof of (20) is presented in Appendix B. If b_i are determined by

$$b_i = \sqrt{(1 - \pi_{i2}) w_i^{-1} u^{-1}}, \quad (21)$$

the variance estimator (16) is consistent because the second term in (20) cancels out \hat{V}_{bias} in (14). This is true even when the first-phase sampling rate n/N is not negligible. To guarantee nonnegative replication weights in (18), we require that b_i in (19) is ≤ 1 . If we set $p_k = 0.5$, then

$$b_i = \sqrt{\frac{4(1 - \pi_{i2}) w_i^{-1}}{\sum_{k=1}^L c_k}},$$

which is less than or equal to 1 if $\sum_{k=1}^L c_k \geq 4$. In fact, the p_k 's can be chosen to be any number between 0 and 1 as long as the resulting b_i in (21) is less than or equal to 1.

4. Extensions

In this section, we consider some extensions of the proposed replication method to types of two-phase estimators other than the REE in (10).

4.1 Double expansion estimator

In two-phase sampling, the double expansion estimator, termed by Kott and Stukel (1997), is also used. The double expansion estimator (DEE) has the simple form

$$\bar{y}_{\text{DEE}} = \frac{1}{N} \sum_{h=1}^H \sum_{i \in A_{h2}} w_i \pi_{i2}^{-1} y_i. \quad (22)$$

When the second-phase sample is a stratified random sample, $\pi_{i2} = r_h/n_h$ and the KNF method can be applied using the replicate

$$\bar{y}_{\text{DEE}}^{(k)} = \frac{1}{N} \sum_{h=1}^H \left(\frac{\sum_{i \in A_{h2}} w_i^{(k)} w_i^{-1}}{\sum_{i \in A_{h2}} w_i^{(k)} w_i^{-1}} \right) \sum_{i \in A_{h2}} w_i^{(k)} y_i.$$

The KNF variance estimator for DEE is consistent when the first-phase sampling rate is negligible. When the first-phase sampling rate is not negligible, we can use the replication method proposed in Section 3. The proposed replication method for the DEE creates replicates,

$$\bar{y}_{DEE}^{*(k)} = \frac{1}{N} \sum_{h=1}^H \sum_{i \in A_{h2}} w_i^{(k)} w_{i2}^{*(k)} y_i, \quad (23)$$

where

$$w_{i2}^{*(k)} = M_{i2}^{(k)} \frac{\sum_{i \in A_{h1}} w_i^{(k)} w_i^{-1}}{\sum_{i \in A_{h2}} w_i^{(k)} w_i^{-1} M_{i2}^{(k)}},$$

and $M_{i2}^{(k)}$ is the replication factor defined in (19). The bias of the replication variance estimator using replicate (23) is negligible if the replicates are constructed to satisfy (21).

If the second-phase sample is an unequal probability sample within each stratum, the replication method such as (23) is not directly applicable. The DEE in (22) is generally less efficient than the REE in (10). Note that the REE in (10) can also be expressed as

$$\bar{y}_{REE} = \frac{1}{N} \sum_{h=1}^H \sum_{i \in A_{h2}} w_i w_{i2}^* y_i, \quad (24)$$

where

$$w_{i2}^* = \pi_{i2}^{-1} \frac{\sum_{i \in A_{h1}} w_i}{\sum_{i \in A_{h2}} w_i \pi_{i2}^{-1}}. \quad (25)$$

The replicates (17) can be written

$$\bar{y}_{REE}^{*(k)} = \frac{1}{N} \sum_{h=1}^H \sum_{i \in A_{h2}} w_i^{(k)} w_{i2}^{*(k)} y_i, \quad (26)$$

where

$$w_{i2}^{*(k)} = M_{i2}^{(k)} \pi_{i2}^{-1} \frac{\sum_{i \in A_{h1}} w_i^{(k)}}{\sum_{i \in A_{h2}} w_i^{(k)} M_{i2}^{(k)} \pi_{i2}^{-1}} \quad (27)$$

and $M_{i2}^{(k)}$ is defined in (19).

4.2 Regression estimator

In two-phase sampling, auxiliary variables that are observed in the first-phase sample can be further used at the estimation stage. The two-phase regression estimator of the population total can be written in the form

$$\hat{Y}_{t, \text{REG}} = \hat{\mathbf{T}}_{x,1}' \hat{\boldsymbol{\beta}}_2 \quad (28)$$

where $\hat{\mathbf{T}}_{x,1} = \sum_{i \in A_1} w_i \mathbf{x}_i$ is the vector of estimated population totals of the control variable \mathbf{x}_i estimated with the first-phase sample and $\hat{\boldsymbol{\beta}}_2 = (\sum_{i \in A_2} w_i w_{i2}^* \mathbf{x}_i \mathbf{x}_i')^{-1} \sum_{i \in A_2} w_i w_{i2}^* \mathbf{x}_i y_i$ is a vector of estimated regression coefficients estimated with the second-phase sample and w_{i2}^* is given by (25). Note that the regression estimator in (28) can incorporate the stratified sampling design in the second-phase if \mathbf{x}_i includes the vector of stratum indicators.

Using the arguments of Section 3, the k^{th} replicate for $\hat{Y}_{t, \text{REG}}$ can be constructed by

$$\hat{Y}_{t, \text{REG}}^{(k)} = \hat{\mathbf{T}}_{x,1}^{(k)'} \hat{\boldsymbol{\beta}}_2^{(k)}, \quad (29)$$

where

$$\hat{\mathbf{T}}_{x,1}^{(k)} = \sum_{i \in A_1} w_i^{(k)} \mathbf{x}_i$$

$$\hat{\boldsymbol{\beta}}_2^{(k)} = \left(\sum_{i \in A_2} w_i^{(k)} w_{i2}^{*(k)} \mathbf{x}_i \mathbf{x}_i' \right)^{-1} \sum_{i \in A_2} w_i^{(k)} w_{i2}^{*(k)} \mathbf{x}_i y_i$$

and $w_{i2}^{*(k)}$ is defined in (27).

The replication method (29) can be directly applicable to the two-phase calibration estimator that was discussed in Hidirolou and Särndal (1998). If $H = 1$, then the replicate of $\hat{\boldsymbol{\beta}}_2$ in (29) reduces to

$$\hat{\boldsymbol{\beta}}_2^{(k)} = \left(\sum_{i \in A_2} w_i^{(k)} M_{i2}^{(k)} \pi_{i2}^{-1} \mathbf{x}_i \mathbf{x}_i' \right)^{-1} \sum_{i \in A_2} w_i^{(k)} M_{i2}^{(k)} \pi_{i2}^{-1} \mathbf{x}_i y_i.$$

5. Simulation study

To study the finite sample performance of the proposed estimators, we conducted a limited simulation study. In the simulation, we first generated an artificial finite population of size $N = 1,000$ with five variables $(z_i, q_i, x_i, y_i, u_i)$, where the population elements are independently generated from $z_i \sim \exp(1) + 2$; $q_i \sim \chi^2(1) + 2$; $x_i \sim N(2, 1)$; $u_i \sim \text{Unif}\{1, 2, 3, 4\}$, where $\text{Unif}\{1, \dots, G\}$ denotes a discrete uniform distribution with support $\{1, \dots, G\}$; and

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 z_i + \beta_3 q_i + e_i$$

with $(\beta_0, \beta_1, \beta_2, \beta_3) = (0, 2, 1, 1)$ and $e_i \sim N(0, 1)$. The variables z_i, q_i, x_i, u_i , and e_i are mutually independent. The stratum for the second-phase sampling was defined using variable u_i . Variable x_i was used to compute the two-phase regression estimator (28) with $\mathbf{x}_i = (1, x_i)'$, variable z_i was used as a size measure for the unequal probability sampling in the first phase sampling, and variable q_i was used as a size measure for the unequal probability sampling in the second phase sampling.

To obtain unequal probability samples for this simulation study, we used either Poisson sampling or Rao-Sampford sampling (Rao 1965 and Sampford 1967), with selection probabilities proportional to the measure of the size variable. Note that the final sample size is random under Poisson sampling but is fixed under Rao-Sampford sampling.

The simulation setup employed a $2 \times 3 \times 2$ factorial structure with three factors. The factors are

1. Sampling for the first-phase sample (2): Simple random sampling of size $n=200$ versus the Rao-Sampford sampling of size $n=200$ using z_i as the measure of size.

2. Sampling for the second-phase sample (3): Stratified random sampling of size $r_h = 25$, stratified Poisson sampling with expected sample size $r_h = 25$ using q_i as the size measure for the unequal probability sampling, and stratified Rao-Sampford sampling of size $r_h = 25$ using q_i as the size measure for the unequal probability sampling.
3. Variance estimation methods (2): The KNF estimator (11) without additional replication versus the proposed variance estimator using (16) were computed based on the jackknife method.

From the finite population generated above, we generated $B = 5,000$ independent Monte Carlo samples for simulation. For the designs with Rao-Sampford sampling in the first phase, we used the jackknife variance estimation method proposed by Berger (2007), which gives a consistent estimator of the first phase sampling variance. The parameter of interest is the population mean of the y variable. From each Monte Carlo sample, we computed two point estimators, the REE in (24) and the regression estimator (REG) in (28) using the auxiliary variable $(1, x_i)$. Relative biases of the variance estimators were computed by dividing the Monte Carlo bias of the variance estimator by the Monte Carlo variance of the point estimator.

Table 1 shows the mean and variance of the two point estimators. For point estimation, the regression estimator is significantly more efficient than the REE for this population because the auxiliary variable x is correlated with the study variable y . The theoretical asymptotic variance of the regression estimator under simple random sampling in the first phase and stratified random sampling in the second phase is approximately equal to

$$\left(\frac{1}{200} - \frac{1}{1,000}\right)8 + \left(\frac{1}{100} - \frac{1}{200}\right)4 = 0.052$$

and the theoretical asymptotic variance of the REE under the same design is, approximately, $(1/100 - 1/1,000)8 = 0.072$, which is consistent with the numerical results in Table 1. The Rao-Sampford sampling in the second phase is slightly more efficient than the Poisson sampling because of the fixed sample size in the Rao-Sampford sampling.

Table 2 shows the relative bias (RB) and coefficient of variation (CV) of the two variance estimators. Relative biases of the variance estimators were computed by dividing the Monte Carlo bias of the variance estimator by the Monte Carlo variance of the point estimator. Coefficients of variation of the variance estimator were computed by dividing the Monte Carlo standard error of the variance estimator by the Monte Carlo average of the variance estimator.

Table 1
Mean and variance of the point estimators (5,000 samples)

Estimator	First-phase Sampling	Second-Phase Sampling	Mean	Variance
REE	SRS	St. SRS	10.0	0.0749
		St. Poi	10.0	0.0784
		St. RS	10.0	0.0754
	RS	St. SRS	10.0	0.0768
		St. Poi	10.0	0.0827
		St. RS	10.0	0.0781
REG	SRS	St. SRS	10.0	0.0540
		St. Poi	10.0	0.0510
		St. RS	10.0	0.0495
	RS	St. SRS	10.0	0.0551
		St. Poi	10.0	0.0531
		St. RS	10.0	0.0515

REE: reweighted expansion estimator (23),
 REG: regression estimator (27),
 SRS: Simple random sampling,
 RS: Rao-Sampford sampling,
 St. SRS: Stratified simple random sampling,
 St. Poi: Stratified Poisson sampling,
 St. RS: Stratified Rao-Sampford sampling.

Table 2
Relative bias (RB) and coefficient of variation (CV) for the variance estimators (5,000 samples)

Method	Estimator	First-phase Sampling	Second-Phase Sampling	RB (%)	CV (%)
KNF	REE	SRS	St. SRS	-11.25	18.22
			St. Poi	-9.56	18.67
			St. RS	-7.75	15.35
		RS	St. SRS	-8.05	18.61
			St. Poi	-9.03	20.84
			St. RS	-5.73	17.27
	REG	SRS	St. SRS	-6.76	22.32
			St. Poi	-6.06	15.81
			St. RS	-3.26	12.82
		RS	St. SRS	-4.17	21.74
			St. Poi	-3.64	16.92
			St. RS	-3.20	13.78
New	REE	SRS	St. SRS	0.09	18.23
			St. Poi	-1.23	19.70
			St. RS	-0.04	16.06
		RS	St. SRS	0.78	19.78
			St. Poi	-2.07	21.26
			St. RS	1.00	17.67
	REG	SRS	St. SRS	-0.61	22.00
			St. Poi	-0.57	16.55
			St. RS	-0.08	13.36
		RS	St. SRS	0.67	22.86
			St. Poi	-0.01	16.97
			St. RS	0.59	14.02

KNF: Kim *et al.* (2006) variance estimator without additional replicates for bias correction,
 New: the proposed variance estimator (16),
 REE: reweighted expansion estimator (23),
 REG: regression estimator (27),
 SRS: Simple random sampling,
 RS: Rao-Sampford sampling,
 St. SRS: Stratified simple random sampling,
 St. Poi: Stratified Poisson sampling,
 St. RS: Stratified Rao-Sampford sampling.

In this simulation, because the first-phase sampling fraction is not negligible ($n/N = 0.2$), the KNF variance estimator without additional replicates underestimates the true variance and the proposed variance estimator estimates the variance with smaller bias, less than 3% in absolute values in all cases, which is consistent with the theory in Section 3 and Section 4. The absolute value of the relative biases in the KNF variance estimator are big because, although in (29) the variance due to $\hat{\mathbf{T}}_{s,1}$ is consistently estimated, the variance due to $\hat{\beta}_2$ is underestimated without additional replicates. The relative biases in our proposed variance estimator are reduced because replicates (18) create additional variation in the replication weights through additional perturbation δ_k drawn from a properly chosen distribution. The proposed variance estimator shows slightly bigger CVs than the KNF method because it involves extra randomness due to generating δ_{ki} from (15).

6. Concluding remarks

Replication variance estimation under two-phase sampling is an importance practical problem in survey sampling and the KNF method is a useful tool in this direction. In this article, we propose an extension of the KNF method in that it can be directly applicable when the first-phase sampling rate is non-negligible, without increasing the number of replicates. The proposed method is also applicable to unequal probability Poisson sampling within each stratum in the second-phase sample. Although the theory has been developed only under Poisson sampling in the second phase, the simulation results in section 5 show that the proposed method works reasonably well for other unequal probability sampling designs, such as the Rao-Sampford sampling design. Since the proposed replication method provides consistent variance estimators for population means, it can be readily applied to other finite population parameters which are smooth functions of population means.

In some large scale surveys, the number of replicates can be quite large because it uses the same number of replicates for the first-phase sample. If one wishes to reduce the number of replicates further, the method of Fuller (1998) or Kim and Sitter (2003) can be considered. Further investigation in this direction will be a topic of future study.

Acknowledgements

The research was supported by a Cooperative Agreement No. 68-3A75-4-122 between the USDA Natural Resources Conservation Service and the Center for Survey Statistics and Methodology at Iowa State University. The authors wish to thank Wayne Fuller and two anonymous referees for helpful comments.

Appendix

A. Proof of (13)

Let $\mathbf{a} = (a_1, \dots, a_N)$ where a_i is the extended version of the second-phase sampling indicator as discussed in Kim *et al.* (2006). That is, $a_i = 1$ if unit i is selected for the second-phase sample once it is in the first-phase sample and $a_i = 0$ otherwise.

By assumption (9), conditional on \mathbf{a} , we have

$$\sum_{k=1}^L c_k (\bar{y}_{h2}^{(k)} - \bar{y}_{h2})^2 = \text{Var}(\bar{y}_{h2} | \mathbf{a}) + o_p(n^{-1}).$$

Thus, the bias of $\sum_{k=1}^L c_k (\bar{y}_{h2}^{(k)} - \bar{y}_{h2})^2$ as an estimator for $\text{Var}(\bar{y}_{h2})$ is then equal to, ignoring $o(n^{-1})$ terms,

$$E\{\text{Var}(\bar{y}_{h2} | \mathbf{a})\} - \text{Var}(\bar{y}_{h2}) = \text{Var}\{E(\bar{y}_{h2} | \mathbf{a})\}.$$

Using the extended definition of a_i , we have

$$E(\bar{y}_{h2} | \mathbf{a}) = \frac{\sum_{i \in U_h} \pi_{i2}^{-1} a_i y_i}{\sum_{i \in U_h} \pi_{i2}^{-2} a_i}$$

and, by the Poisson sampling assumption of a_i 's,

$$\text{Var}\left(\frac{\sum_{i \in U_h} \pi_{i2}^{-1} a_i y_i}{\sum_{i \in U_h} \pi_{i2}^{-2} a_i}\right) = N_h^{-2} \sum_{i \in U_h} \pi_{i2}^{-1} (1 - \pi_{i2}) (y_i - \bar{y}_h)^2 + o(N^{-1}). \quad (\text{A.1})$$

Thus, the bias of the KNF variance estimator is of the form (13) under the Poisson sampling assumption of a_i .

B. Proof of (20)

For each k ,

$$\bar{y}_{ip}^{*(k)} - \bar{y}_{ip} = \bar{y}_{ip}^{*(k)} - \bar{y}_{ip}^{(k)} + \bar{y}_{ip}^{(k)} - \bar{y}_{ip},$$

where $\bar{y}_{ip}^{(k)}$ is defined in (12). Thus,

$$\begin{aligned} \hat{V}_{\text{KNF}}^* &= \sum_{k=1}^L c_k (\bar{y}_{ip}^{*(k)} - \bar{y}_{ip})^2 = \sum_{k=1}^L c_k (\bar{y}_{ip}^{(k)} - \bar{y}_{ip})^2 \\ &\quad + 2 \sum_{k=1}^L c_k (\bar{y}_{ip}^{(k)} - \bar{y}_{ip}) (\bar{y}_{ip}^{*(k)} - \bar{y}_{ip}^{(k)}) \\ &\quad + \sum_{k=1}^L c_k (\bar{y}_{ip}^{*(k)} - \bar{y}_{ip}^{(k)})^2. \end{aligned} \quad (\text{B.1})$$

By the construction of $\bar{y}_{ip}^{*(k)}$, we have

$$E_*(\bar{y}_{ip}^{*(k)}) = \bar{y}_{ip}^{(k)} + o_p(n^{-1}). \quad (\text{B.2})$$

Also, writing $q_{ki} = M_{i2}^{(k)} - 1$, we have $q_{ki} = O_p(n^{-1/2})$ and we can apply a Taylor expansion to get

$$\bar{y}_{h2}^{*(k)} = \bar{y}_{h2}^{(k)} + \frac{\sum_{i \in A_{h2}} w_i^{(k)} \pi_{i2}^{-1} q_{ki} (y_i - \bar{y}_{h2}^{(k)})}{\sum_{i \in A_{h2}} w_i^{(k)} \pi_{i2}^{-1}} + o_p(n^{-1}). \quad (\text{B.3})$$

Also, because

$$\frac{1}{N_h} \sum_{i \in A_{h2}} w_i^{(k)} \pi_{i2}^{-1} z_i - \frac{1}{N_h} \sum_{i \in A_{h2}} w_i \pi_{i2}^{-1} z_i = O_p(n^{-1})$$

for any z variable with bounded fourth moments, it can be shown that (B.3) reduces to

$$\bar{y}_{h2}^{*(k)} = \bar{y}_{h2}^{(k)} + \frac{\sum_{i \in A_{h2}} w_i \pi_{i2}^{-1} q_{ki} (y_i - \bar{y}_{h2})}{\sum_{i \in A_{h2}} w_i \pi_{i2}^{-1}} + o_p(n^{-1}).$$

Hence, we can write

$$\sum_{k=1}^L c_k (\bar{y}_{lp}^{*(k)} - \bar{y}_{lp}^{(k)})^2 = \sum_{k=1}^L c_k \left\{ N^{-1} \sum_{h=1}^H \sum_{i \in A_{h2}} w_i \pi_{i2}^{-1} q_{ki} (y_i - \bar{y}_{h2}) \right\}^2 + o_p(n^{-1}). \quad (\text{B.4})$$

Inserting (B.2) and (B.4) into (B.1), we have

$$\begin{aligned} E_*(\hat{V}_{\text{KNF}}^*) &= \hat{V}_{\text{KNF}} \\ &+ \frac{1}{N^2} \sum_{k=1}^L c_k \sum_{h=1}^H \sum_{i \in A_{h2}} w_i^2 E_*(q_{ki}^2) \pi_{i2}^{-2} (y_i - \bar{y}_{h2})^2 \\ &+ o_p(n^{-1}), \end{aligned}$$

and because $E_*(q_{ki}^2) = p_k(1 - p_k) b_i^2$, we have (20).

References

- Berger, Y.G. (2007). A jackknife variance estimator for unistage stratified samples with unequal probabilities. *Biometrika*, 94, 953-964.
- Binder, D.A., Babyak, C., Brodeur, M., Hidirolou, M. and Jocelyn, W. (2000). Variance estimation for two-phase stratified sampling. *The Canadian Journal of Statistics*, 28, 751-764.
- Breidt, F.J. and Fuller, W.A. (1993). Regression weighting for multipurpose samplings. *Sankhyā*, B, 55, 297-309.
- Brick, J.M., and Morganstein, D. (1996). WesVarPC: Software for computing variance estimates from complex designs. *Proceedings of the 1996 Annual Research Conference*, U.S. Bureau of the Census, 861-866.
- Cochran, W.G. (1977). *Sampling Techniques*. New York: John Wiley & Sons, Inc.
- Fay, R.E. (1984). Some properties of estimates of variance based on replication methods. *Proceedings of the Survey Research Method Section*, American Statistical Association, 495-500.
- Flyer, P. (1987). Finite population correction for replication estimates of variance. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 732-736.
- Fuller, W.A. (1998). Replication variance estimation for two-phase samples. *Statistica Sinica*, 8, 1153-1164.
- Fuller, W.A. (2003). Estimation for multiple phase samples. In *Analysis of Survey Data*, (Eds., R.L. Chambers and C.J. Skinner). Wiley, Chichester, England, 307-322.
- Hidirolou, M.A. (2001). Double sampling. *Survey Methodology*, 27, 143-154.
- Hidirolou, M.A., and Särndal, C.-E. (1998). Use of auxiliary information for two-phase sampling. *Survey Methodology*, 24, 11-20.
- Kim, J.K., Navarro, A. and Fuller, W.A. (2006). Replicate variance estimation after multi-phase stratified sampling. *Journal of the American Statistical Association*, 101, 312-320.
- Kim, J.K., and Sitter, R.R. (2003). Efficient variance estimation for two-phase sampling. *Statistica Sinica*, 13, 641-653.
- Kott, P.S., and Stukel, D.M. (1997). Can the jackknife be used with a two-phase Sample? *Survey Methodology*, 23, 81-89.
- Lu, W., and Sitter, R.R. (2006). Disclosure risk and variance estimation. *Proceedings of Statistics Canada international symposium series*, 11-522-XIE.
- Neyman, J. (1938). Contribution to the theory of sampling human populations. *Journal of the American Statistical Association*, 33, 101-116.
- Rao, J.N.K. (1965). On two simple schemes of unequal probability sampling without replacement. *Journal of the Indian Statistical Association*, 3, 173-180.
- Rao, J.N.K. (1973). On double sampling for stratification and analytical surveys. *Biometrika*, 60, 125-133.
- Rao, J.N.K., and Shao, J. (1992). Jackknife variance estimation with survey data under hot deck imputation. *Biometrika*, 79, 811-822.
- Rao, J.N.K., and Sitter, R.R. (1995). Variance estimation under two-phase sampling with application to imputation for missing data. *Biometrika*, 82, 453-460.
- Rust, K.F., and Rao, J.N.K. (1996). Variance estimation for complex surveys using replication techniques. *Statistical Methods in Medical Research*, 5, 283-310.
- Sampford, M.R. (1967). On sampling without replacement with unequal probability of selection. *Biometrika*, 54, 499-513.
- Wolter, K. (2007). *Introduction to Variance Estimation*. 2nd Edition, New York: Springer.

Cost efficiency of repeated cluster surveys

Stanislav Kolenikov and Gustavo Angeles¹

Abstract

We analyze the statistical and economic efficiency of different designs of cluster surveys collected in two consecutive time periods, or waves. In an independent design, two cluster samples in two waves are taken independently from one another. In a cluster-panel design, the same clusters are used in both waves, but samples within clusters are taken independently in two time periods. In an observation-panel design, both clusters and observations are retained from one wave of data collection to another. By assuming a simple population structure, we derive design variances and costs of the surveys conducted according to these designs. We first consider a situation in which the interest lies in estimation of the change in the population mean between two time periods, and derive the optimal sample allocations for the three designs of interest. We then propose the utility maximization framework borrowed from microeconomics to illustrate a possible approach to the choice of the design that strives to optimize several variances simultaneously. Incorporating the contemporaneous means and their variances tends to shift the preferences from observation-panel towards simpler panel-cluster and independent designs if the panel mode of data collection is too expensive. We present numeric illustrations demonstrating how a survey designer may want to choose the efficient design given the population parameters and data collection cost.

Key Words: Longitudinal study; Cluster samples; DHS; NHIS.

1. Introduction

To analyze the dynamics of social, behavioral or population health phenomena, researchers and policymakers need to obtain information on characteristics of the population on multiple occasions. Complex design surveys are the most frequently used sources of information for large populations, such as a country as a whole. Besides the standard considerations in single-shot surveys, *e.g.*, stratification and clustering, other issues may be important in surveys collected over two or more time periods. In such surveys, the total cost and the total survey error are affected by an overlap among consecutive samples, (informative) sample attrition, time-in-sample or conditioning effects, and other dynamic factors.

For the purposes of estimation of change from repeated surveys, it is often desirable to have high temporal correlation of the observation units which can be achieved by administering the survey to the same sampling and/or observation units. In longitudinal surveys, the same observation units (individuals, households) are revisited for several periods, potentially indefinitely many periods (the US Panel Study of Income Dynamics (PSID), British Household Panel Study (BHPS) and others). A compendium of information on the longitudinal studies can be found at the Institute for Social and Economics Research web site, <http://iser.essex.ac.uk/ulsc/keeptrack/index.php>). In rotating panel surveys, the observation units are recruited into the sample for a few periods, then rotated out of the sample, and surveyed again at a later time. Examples of rotating panel

surveys include the US Current Population Survey (CPS) (Binder and Hidioglou 1988, Eckler 1955, Rao and Graham 1964) and a number of environmental surveys (Fuller 1999, McDonald 2003, Scott 1998). Yet another option is to use the same primary sampling units (PSUs) in different waves, but sample the observation units (secondary sampling units, SSUs) independently. Surveys collected in this way include international Demographic and Health Surveys (DHS) and the US National Health Interview Survey (NHIS).

We shall concentrate on surveys collected in two time periods, or waves, using a two-stage cluster design in each wave of data collection. We consider three possible designs differing in the amount and depth of overlap of sampling units over time. The sample designer can simply ignore any possible effects arising from the sample overlap, and take two independent samples in two periods of time. We shall refer to this design as the *independent* design. Alternatively, the sample designer may find it beneficial to recycle the PSUs from one wave to another. If the designer finds it difficult to track the SSUs from one wave to another, the subsamples within clusters can be taken independently in two waves of data collection. We shall refer to this design as the *cluster-panel* design. If an utmost precision is essential, the fully longitudinal design will attempt to locate all individuals who responded in the first wave, and solicit the second interview. To distinguish this design from the cluster-panel design, we shall refer to it as the *observation-panel* design.

1. Stanislav Kolenikov, Department of Statistics, 146 Middlebush Hall, University of Missouri, Columbia, MO 65211-6100, U.S.A. E-mail: kolenikovs@missouri.edu; Gustavo Angeles, Associate Director of the Center for Evaluation Research, National Institute of Public Health, Mexico, Mexico. E-mail: gangeles@insp.mx.

A particular aspect that we found important in survey management, but underaddressed in the existing literature, is the implementation cost (Groves 1989). The traditional cost models such as those used in derivation of Neyman-Tchuprow optimal allocation design (Neyman 1938) can be extended to include terms related to the cost of the first visit to the cluster and ultimate observation unit, as well as the cost of consecutive visits. The cost of revisiting the cluster is likely to be lower on the second occasion. There is no need to create new maps and set up frames. The same interviewers can be used to conduct interviews in subsequent waves of data collection. Cooperation with community leaders has been established earlier, if it is important, as it is in some traditional societies. The effect of the panel mode of data collection at the individual level is less clear. If the household that was interviewed in earlier waves moved out and would have to be located, possibly in different geographic area, the (average) cost of the panel interview goes up. The likelihood of such circumstances increases with longer intervals between surveys typical for the developing countries surveys: the intervals between waves of DHS are usually about 5-7 years. On the other hand, if a less expensive interview mode can be used after the first round, (e.g., a phone interview instead of the personal visit), the cost of the panel interview goes down.

This paper brings together statistical and economic considerations in the choice of the appropriate design and its parameters. We assume the survey designer can be interested in estimating the change in the population mean between two time periods, and/or the means themselves. We introduce a sketchy population in Section 2, and compute the design variances of the means and their differences for the three sampling designs of our interest.

To incorporate economic aspects of data collection, we introduce a relatively simple cost model for a repeated cluster survey in Section 3. We set up and solve optimization problems to obtain the optimal sample sizes for the three considered designs. By plugging in the estimates of the statistical parameters (variances and autocorrelations) and cost components (cluster-level and individual-level costs), the survey designer can compare the numeric values of the variances to choose the best design. Section 4 illustrates this approach and shows that each of the designs may be the best one, depending on the parameter values. The intuitive results (e.g., the higher cost of data collection and lower autocorrelations of the observed characteristics make panel modes of data collection less appealing) are given an analytic justification and quantitative backing.

While Sections 2-4 deal with the efficiency in estimating the difference in means only, more realistic goals of data collection efforts would include contemporaneous characteristics and their variances. To this end, Section 5

introduces a utility maximization framework describing the survey designer's choice of the sampling scheme. This framework provides an aggregated objective function that combines several design criteria. The results are again as expected: if the more expensive panel modes of data collection result in smaller sample sizes, the estimates of the means are less efficient than in simpler designs. The only way to justify these efficiency losses is by a drastic improvement in the estimation of the difference that can only occur with higher autocorrelations. Such effects are also illustrated in Section 5. Section 7 concludes. Proofs are given in the Appendix.

2. Design variances

Let the population consist of N clusters, or PSUs, in both time periods, and each cluster consist of M individuals, or SSUs. Out of these, an SRS of $1 < n_t \leq N$ clusters is taken at time $t = 1, 2$, and an SRS of $1 < m_t \leq M$ individuals is taken in each cluster that is present in the sample at time t . Let the index i denote PSUs, and the index j , SSUs. Thus the typical measurement will be denoted as Y_{ij} in the population, and y_{ij} in the sample. The population totals $T[\cdot]$ and their estimates $t[\cdot]$ can then be found as follows:

cluster total:

$$T_{i\cdot}[Y] = \sum_{j=1}^M Y_{ij}, \quad t_{i\cdot}[y] = \frac{M}{m} \sum_{j=1}^M y_{ij},$$

population total:

$$T_t[Y] = \sum_{i=1}^N Y_{i\cdot}, \quad t_t[y] = \frac{N}{n} \sum_{i=1}^N t_{i\cdot}[y]. \quad (2.1)$$

The means per observation units are

$$\bar{Y}_{i\cdot} = \frac{1}{M} \sum_{j=1}^M Y_{ij} = \frac{T_{i\cdot}[Y]}{T_{i\cdot}[1]}, \quad \bar{y}_{i\cdot} = \frac{1}{m} \sum_{j=1}^m y_{ij} = \frac{t_{i\cdot}[y]}{t_{i\cdot}[1]},$$

$$\bar{Y}_{t\cdot} = \frac{T_t[Y]}{T_t[1]} = \frac{\sum_{i=1}^N \sum_{j=1}^M Y_{ij}}{NM}, \quad \bar{y}_{t\cdot} = \frac{t_t[y]}{t_t[1]} = \frac{\sum_{i=1}^n \sum_{j=1}^m y_{ij}}{nm}. \quad (2.2)$$

The variance of Y and its within- and between-cluster components are

$$S_t^2 = \frac{\sum_{i=1}^n \sum_{j=1}^M (Y_{ij} - \bar{Y}_{i\cdot})^2}{NM - 1}, \quad (2.3)$$

$$S_{tw}^2 = \frac{\sum_{i=1}^n (Y_{i\cdot} - \bar{Y}_{t\cdot})^2}{M - 1}, \quad \bar{S}_{tw}^2 = \frac{1}{N} \sum_{i=1}^N S_{twi}^2, \quad (2.4)$$

$$S_{ib}^2 = \frac{\sum_{i=1}^N (\bar{Y}_{i.} - \bar{Y}_{.})^2}{N-1}. \quad (2.5)$$

The characteristic of primary interest is the change in the means,

$$D = \bar{Y}_{2..} - \bar{Y}_{1..}, \quad (2.6)$$

estimated by

$$d = \bar{y}_{2..} - \bar{y}_{1..}. \quad (2.7)$$

An attractive property of this estimator for analysts and data users is its internal consistency: the estimator of the difference is the difference of the estimators. If the samples in consecutive periods overlap only partially, then composite or GLS estimators (Fuller 1999, Hansen, Hurwitz and Madow 1953, Patterson 1950, Rao and Graham 1964, Wolter 2007) have better efficiency.

In what follows, we assume all sampling procedures to be simple random sampling without replacement. For the contemporaneous mean, the variance is given by (Cochran 1977, Th. 10.1):

$$V[\bar{y}_{t..}] = \left(1 - \frac{n}{N}\right) \frac{S_{tb}^2}{n} + \left(1 - \frac{m}{M}\right) \frac{\bar{S}_{tw}^2}{nm}. \quad (2.8)$$

For simplicity and clarity of exposition, we shall often be making an assumption of symmetric conditions:

$$S_{1wi}^2 = S_{2wi}^2 = S_{wi}^2, \bar{S}_{1w}^2 = \bar{S}_{2w}^2 = \bar{S}_w^2, S_{1b}^2 = S_{2b}^2 = S_b^2. \quad (2.9)$$

Analytic derivations are possible without these assumptions, but become extremely cumbersome. Besides, it is unrealistic to think that the survey designer could know the characteristics of the future population. Thus (2.9) should be viewed as a reasonable working model.

2.1 Independent design

Proposition 1. Let n_1 out of N clusters and m_1 out of M observation units in selected clusters be taken without replacement at time $t = 1$. Let n_2 out of N clusters and m_2 out of M observation units in selected clusters be taken without replacement at time $t = 2$, with sampling performed independently from that at time $t = 1$. Then

$$V_i(d) = \left(1 - \frac{n_1}{N}\right) \frac{S_{1b}^2}{n_1} + \left(1 - \frac{n_2}{N}\right) \frac{S_{2b}^2}{n_2} + \left(1 - \frac{m_1}{M}\right) \frac{\bar{S}_{1w}^2}{n_1 m_1} + \left(1 - \frac{m_2}{M}\right) \frac{\bar{S}_{2w}^2}{n_2 m_2}. \quad (2.10)$$

The result follows immediately from (2.8) by independence of the two samples. The subindex of the variance i stands for the “independent design”. Under the symmetric

conditions of (2.9), if the sample sizes are the same in two periods, $n_1 = n_2 = n$ and $m_1 = m_2 = m$, then

$$V_{e,i}[d] = 2\left(1 - \frac{n}{N}\right) \frac{S_b^2}{n} + 2\left(1 - \frac{m}{M}\right) \frac{\bar{S}_w^2}{nm}, \quad (2.11)$$

where the subindex e, i stands for “equal variances, independent design”.

2.2 Cluster-panel design

Proposition 2. Let n out of N clusters be sampled without replacement in the first period and be used in both time periods. Let m out of M observation units be sampled without replacement independently in two periods. Then

$$V_c[d] = \left(1 - \frac{n}{N}\right) \frac{S_{1b}^2 + S_{2b}^2 - 2\rho^I S_{1b} S_{2b}}{n} + \left(1 - \frac{m}{M}\right) \frac{\bar{S}_{1w}^2 + \bar{S}_{2w}^2}{nm},$$

$$\rho^I = \frac{1}{S_{1b} S_{2b} (N-1)} \sum_{i=1}^N (\bar{Y}_{1i.} - \bar{Y}_{1..})(\bar{Y}_{2i.} - \bar{Y}_{2..}). \quad (2.12)$$

Here, subindex c stands for the “cluster-panel design”, and ρ^I is the intertemporal correlation, or autocorrelation, of the cluster means. The superscript I denotes the first stage of sampling. If ρ^I is positive, then the cluster-panel design is more efficient than the independent design for fixed values of n and m . Under the symmetry conditions,

$$V_{e,c}[d] = 2\left(1 - \frac{n}{N}\right) \frac{S_b^2(1 - \rho^I)}{n} + 2\left(1 - \frac{m}{M}\right) \frac{\bar{S}_w^2}{nm}, \quad (2.13)$$

where the subindex e, c stands for the “equal variances, cluster-panel design”.

2.3 Observation-panel design

Proposition 3. Let n out of N clusters and m out of M observation units be sampled without replacement in the first period and be used in both time periods. Then

$$V_o[d] = \left(1 - \frac{n}{N}\right) \frac{S_{1b}^2 + S_{2b}^2 - 2\rho^I S_{1b} S_{2b}}{n} + \left(1 - \frac{m}{M}\right) \frac{\bar{S}_{1w}^2 + \bar{S}_{2w}^2 - 2\rho^{II} \bar{S}_{1w} \bar{S}_{2w}}{nm},$$

$$\rho^{II} = \frac{1}{\bar{S}_{1w} \bar{S}_{2w} N(M-1)} \sum_{i=1}^N \sum_{j=1}^M (Y_{ij} - \bar{Y}_{i.})(Y_{2ij} - \bar{Y}_{2i.}). \quad (2.14)$$

Subindex o stands for the “observation-panel design”. Under the assumption of symmetric conditions,

$$V_{c,o}[d] = 2 \left(1 - \frac{n}{N} \right) \frac{(1 - \rho^I) S_h^2}{n} + 2 \left(1 - \frac{m}{M} \right) \frac{(1 - \rho^{II}) \bar{S}_n^2}{nm},$$

$$\rho^{II} = \frac{1}{\bar{S}_n^2 N (M-1)} \sum_{i=1}^N \sum_{j=1}^M (Y_{1ij} - \bar{Y}_{1i})(Y_{2ij} - \bar{Y}_{2i}) \quad (2.15)$$

with corresponding e, o subindex for the “equal variances, observation-panel design”.

Here, ρ^{II} is the intertemporal correlation, or autocorrelation, of the individual observations within clusters. The superscript II stands for the second stage of sampling. If ρ^{II} is positive, then the observation-panel design is more efficient than the cluster-panel design for fixed values of n and m .

How are the two autocorrelations that appear in (2.15) related? Conceptually, one can think of any number of possible relations between them. Let us introduce a superpopulation model

$$Y_{ij} = \mu_t + a_{it} + \varepsilon_{ij}, \quad E_{\xi}[a_{it}] = 0, \quad E_{\xi}[\varepsilon_{ij}] = 0, \quad (2.16)$$

in which a_{it} and ε_{ij} are independent of one another for all $s, t = 1, 2$. The subindex ξ stands for the superpopulation model expectations. The case of $\rho^I = 0$ and $\rho^{II} = 1$ occurs when the changes in the cluster means occur independently between clusters ($E_{\xi}[a_{1i}a_{2i}] = 0$), but the individuals retain their positions within the cluster, $\varepsilon_{1ij} = \varepsilon_{2ij}$. The case of $\rho^I = 1$ and $\rho^{II} = 0$ occurs when the cluster random effects are the same in both periods, $a_{1i} = a_{2i}$, while the individual random effects are uncorrelated ($E_{\xi}[\varepsilon_{1ij}\varepsilon_{2ij}] = 0$). Neither of these situations is entirely realistic. However, it can probably be expected that the individual, rather than the cluster, dynamics are a more important source of variation over time, thus making the relations $\rho^{II} \geq \rho^I \geq 0$ the most plausible ones. We shall study in numeric examples of Sections 4 and 5 the extent to which the choice of the best design is sensitive to the relation between the two correlations.

3. Costs for repeated cluster samples

In this section we shall analyze the cost efficiency of cluster samples when one wants to estimate the difference between two sample means from two different periods.

Some discussion of the costs of cluster sampling is given in Kish (1995, Section 8.3B), Thompson (1992, Section 12.5), and Lehtonen and Pahkinen (2004). More mathematical details are available in Hansen *et al.* (1953, volume II, Section 6.11), with the variance formulas corrected for finite populations.

3.1 Notation and cost models

Let us assume the following cost structure, which is an extension of Kish (1995) for repeated surveys:

- c_1^I is the cluster level cost at time $t = 1$ for clusters that are used *in the first wave only*;
- c_2^I is the cluster level cost for a *new* cluster at time $t = 2$;
- c_{12}^I is the cluster level cost for clusters in which the data are collected in both periods $t = 1$ and $t = 2$ (PSU panel cost);
- c_1^{II} is the individual level cost at time $t = 1$ for individuals that are observed *in the first wave only*;
- c_2^{II} is the individual level cost at time $t = 2$ for individuals that are observed *in the second wave only*;
- c_{12}^{II} is the individual level cost if the unit is observed in both periods in the observation-panel design (SSU panel cost);
- C_0 is the total budget allocated to the field work in both time periods.

Roman superscripts denote the sampling stage. Arabic subscripts correspond to the occasion at which the sample is taken. The cluster level costs include the cost of sampling the clusters, obtaining the PSU maps, collecting community data, local interviewer training, *etc.* The individual level costs are mostly those of the personal interviews with the ultimate observation units. The total cost C_0 is thought of as the variable cost of the survey that is directly related to the number of sampled units. Fixed cost, such as the cost of preparing the survey instrument and other organization-level costs are not part of C_0 .

3.2 Independent design

The budget constraint for the independent design is given by

$$C_0 = c_1^I n_1 + c_1^{II} n_1 m_1 + c_2^I n_2 + c_2^{II} n_2 m_2. \quad (3.1)$$

The first two terms are the costs of the first wave of data collection, and the last two terms, of the second wave.

Proposition 4. If the survey setting parameters are the same in the two time periods:

$$c_1^I = c_2^I = c^I, \quad c_1^{II} = c_2^{II} = c^{II}, \quad (3.2)$$

then the optimal sample sizes and the resulting variances are given by

$$\begin{aligned}
m &= \sqrt{\frac{c^I \bar{S}_w^2}{c^{II} S_b^2 - \bar{S}_w^2 / M}}, \\
n &= \frac{C_0}{2 \{c^I + [c^I c^{II} \bar{S}_w^2 / (S_b^2 - \bar{S}_w^2 / M)]^{1/2}\}}, \\
V_{e,c}[d] &= \frac{4 \left[c^I + \sqrt{c^I c^{II} \bar{S}_w^2 / (S_b^2 - \bar{S}_w^2 / M)} \right]}{C_0} \\
&\times \left[S_b^2 + \left(\sqrt{\frac{c^{II} S_b^2 - \bar{S}_w^2 / M}{c^I \bar{S}_w^2}} - \frac{1}{M} \right) \bar{S}_w^2 - \frac{2}{N} S_b^2 \right]. \quad (3.3)
\end{aligned}$$

In equations (3.3), the sample sizes n and m are treated as continuous variables. In practice, the nearest integer should be used, with a minimum of 2 necessary to estimate the appropriate variance component, and the maxima of N and M , respectively.

The number of observations sampled within a cluster depends only on the relative costs at the cluster and the observation level, c^I/c^{II} , and relative variances S_b^2/\bar{S}_w^2 , or equivalently the intraclass correlation. Greater interview cost c^{II} prevents the sample designer from using more observations: an increase in c^{II} leads to a decrease in both m and n . Greater cluster-level cost leads to redistribution of the sampled units: n decreases with c^I , while m increases with it. Greater within-cluster variance \bar{S}_w^2 necessitates a greater number of observations m to be taken within a cluster to maintain overall precision. Greater between-cluster variance S_b^2 necessitates a greater number of clusters n to be sampled. Finally, the total survey budget C_0 affects the number of clusters n , but not the subsample size m . As a result, the variance of d is inversely proportional to C_0 .

The non-symmetric situation can be treated as a by-product of the first order conditions derived in the proof (see Appendix). However, no analytic solution is available in that case.

3.3 Cluster-panel design

The budget constraint for the cluster-panel design is given by

$$C_0 = c_{12}^I n + c_1^{II} n m_1 + c_2^{II} n m_2. \quad (3.4)$$

The first term is the cluster-level cost associated with the sample design, and the remaining two terms are the costs of collecting individual-level data in the first and the second waves, respectively.

Proposition 5. The sample sizes for the cluster-panel design are given by

$$\begin{aligned}
m_1 &= 2C_0 / c_1^{II} \left(1 + \frac{\bar{S}_{2w}^2}{\kappa \bar{S}_{1w}^2} + \sqrt{D} \right), \\
m_2 &= \kappa m_1, \\
n &= \frac{C_0}{c_{12}^I + c_1^{II} m_1 + c_2^{II} m_2}, \\
\kappa &= \sqrt{\frac{c_1^{II} \bar{S}_{2w}^2}{c_2^{II} \bar{S}_{1w}^2}}, \quad (3.5)
\end{aligned}$$

provided that

$$\begin{aligned}
D &= \left(1 + \frac{\bar{S}_{2w}^2}{\kappa \bar{S}_{1w}^2} \right)^2 + 8 \frac{(1 - \rho^I) S_b^2 C_0}{\bar{S}_{1w}^2 c_1^{II}} \\
&\quad - 4 \frac{C_0}{c_1^{II} M} \left(1 + \frac{\bar{S}_{2w}^2}{\kappa \bar{S}_{1w}^2} \right) \geq 0.
\end{aligned}$$

The variance of the difference estimator is found by plugging these expressions into (2.13). Under the assumptions of symmetric conditions in two rounds of the survey (2.9) and (3.2),

$$\begin{aligned}
D &= 4 - 8 \frac{C_0}{M c^{II}} + 8 \frac{(1 - \rho^I) S_b^2 C_0}{\bar{S}_w^2 c^{II}}, \\
m_1 &= m_2 = m \\
&= \frac{C_0}{c^{II} + \sqrt{(c^{II})^2 - \frac{2c^{II} C_0}{M} + \frac{2(1 - \rho^I) S_b^2 C_0 c^{II}}{\bar{S}_w^2}}}, \\
n &= \frac{C_0}{c_{12}^I + 2c^{II} m} \\
&= \frac{C_0}{c_{12}^I + 2C_0 / \left[1 + \sqrt{1 - \frac{2c^{II} C_0}{M c^{II}} + \frac{2(1 - \rho^I) S_b^2 C_0}{\bar{S}_w^2 c^{II}}} \right]},
\end{aligned}$$

and $V_{e,c}[d]$ can be found from (2.13).

Interestingly, the number of the SSUs depends on the SSU costs c^{II} , but not on the PSU costs c_{12}^I . An increase in the intraclass correlation, or increase in S_b^2 , or decrease in \bar{S}_w^2 , predictably leads to decrease in the optimal number of SSUs and increase in the optimal number of PSUs. The dependence of the design parameters on the survey budget C_0 is non-trivial. For very small surveys, the number of units per cluster is proportional to C_0 , and the number of clusters is not affected by C_0 . Indeed, if the characteristic demonstrates strong correlation between time periods, it would be preferable to get accurate estimates of the cluster means, and good accuracy of the overall difference estimator will follow. To put it differently, the first term in (2.13) is relatively small by virtue of the positive correlation coefficient ρ^I , and the second term is inversely proportional

to C_0 . For large surveys, $D \propto C_0$, so both the number of units per cluster and the number of clusters are proportional to $\sqrt{C_0}$. The first term in (2.13) is then inversely proportional to $\sqrt{C_0}$, and the second term is inversely proportional to C_0 . An increase in the budget of the survey will affect all terms, although to a different extent.

3.4 Observation-panel design

The budget constraint for the observation-panel design is given by

$$C_0 = c_{12}^I n + c_{12}^{II} nm. \quad (3.6)$$

The first term is the cluster-level cost, and the second term is the cost of individual interviews.

Proposition 6. The optimal sample sizes for the observation-panel design are given by

$$m = \sqrt{\frac{c_{12}^I (1 - \rho^{II}) \bar{S}_w^2}{c_{12}^{II} (1 - \rho^I) S_b^2 - (1 - \rho^{II}) \bar{S}_w^2 / M}},$$

$$n = \frac{C_0}{c_{12}^I + \sqrt{\frac{(1 - \rho^{II}) \bar{S}_w^2 c_{12}^I c_{12}^{II}}{(1 - \rho^I) S_b^2 - (1 - \rho^{II}) \bar{S}_w^2 / M}}}. \quad (3.7)$$

The design variance of the resulting difference estimator is

$$V_{c,o}[d] = \frac{2}{C_0} \left\{ (1 - \rho^I) S_b^2 c_{12}^I \right.$$

$$+ (1 - \rho^{II}) \bar{S}_w^2 \sqrt{\frac{c_{12}^I c_{12}^{II} (1 - \rho^I) S_b^2 - (1 - \rho^{II}) \bar{S}_w^2 / M}{(1 - \rho^{II}) \bar{S}_w^2}}$$

$$+ \left[(1 - \rho^I) S_b^2 - \frac{1}{M} (1 - \rho^{II}) \bar{S}_w^2 \right]$$

$$\times \sqrt{\frac{(1 - \rho^{II}) \bar{S}_w^2 c_{12}^I c_{12}^{II}}{(1 - \rho^I) S_b^2 - (1 - \rho^{II}) \bar{S}_w^2 / M}}$$

$$\left. + (1 - \rho^{II}) \bar{S}_w^2 \left(c_{12}^{II} - \frac{c_{12}^I}{M} \right) \right\} - \frac{2(1 - \rho^I) S_b^2}{N}. \quad (3.8)$$

The sample size expressions (3.7) resemble the ones for the independent design, equation (3.3), with the cost of data collection in a single wave replaced by the cost of panel data collection, and the variance components S_b^2 and \bar{S}_w^2 replaced by $(1 - \rho^I) S_b^2$ and $(1 - \rho^{II}) \bar{S}_w^2$. The second stage sampling size m only depends on the relative cost at the cluster and observation levels, and on the ratio of the variance components augmented by the autocorrelations. Hence, like in the independent design, the dependency of the sample size on the scale of the survey is only through

$n \propto C_0$, and the variance of the difference decreases inversely proportional to C_0 .

Extending the relations between the functional forms of equations (3.3) and (3.8), we can establish the general relations between the two designs:

Proposition 7. If $M \gg 1$ and $N \gg 1$, then $V_{e,i}[d] \geq V_{e,o}[d]$ if

$$2 \left(\sqrt{c^I S_b^2} + \sqrt{c^{II} \bar{S}_w^2} \right)^2$$

$$\geq \left[\sqrt{c_{12}^I (1 - \rho^I) S_b^2} + \sqrt{c_{12}^{II} (1 - \rho^{II}) \bar{S}_w^2} \right]^2. \quad (3.9)$$

Unfortunately, the variance for the cluster-panel design that can be obtained by combining the results of Proposition 5 with (2.13), does not permit an equally lucid comparison.

4. Numeric illustration

To illustrate how the characteristics of population (variances and autocorrelations) and the data collection process (costs) affect the choice of the most efficient design, we consider a numeric example. Let us choose the basic setup with symmetric conditions, and let the parameter values be:

$$N = 10,000, \quad M = 1,000, \quad S_b = 100,$$

$$S_w = 400, \quad \rho^I = 0.1, \quad \rho^{II} = 0.35,$$

$$c_1^{II} = c_2^{II} = 1, \quad c_{12}^{II} = 3, \quad c_1^I = c_2^I = 10,$$

$$c_{12}^I = 18, \quad C_0 = 20,000. \quad (4.1)$$

The cost structure implies that the cost of collecting the initial information for a cluster is the cost of ten interviews, while the cost of the followup in the same cluster is only eight interviews. On the other hand, getting the second interview with the same unit is twice as expensive as getting the first interview.

With these parameters, the sample sizes and design variances are:

$$m_{e,i} = 12, \quad m_{e,c} = 12, \quad m_{e,o} = 8,$$

$$n_{e,i} = 455, \quad n_{e,c} = 476, \quad n_{e,o} = 476,$$

$$m_{e,i} n_{e,i} = 5,460, \quad m_{e,c} n_{e,c} = 5,712, \quad m_{e,o} n_{e,o} = 3,808,$$

$$V_{e,i}[d] = 99.86, \quad V_{e,c}[d] = 91.37, \quad V_{e,o}[d] = 90.20. \quad (4.2)$$

The observation-panel design is 1.2% more efficient than the cluster-panel design, and 10.7% more efficient than the independent design. However, it has a notably smaller total sample size, only 2/3 of the cluster-panel design sample size and 70% of the independent design sample size.

Of course these findings are highly specific to the parameters of the population and the cost structure. Can we describe general patterns of how the variances, and hence the relative efficiency of different designs, change with those parameters? The variances in (4.2) are derived from 13 parameters given in (4.1), and it is difficult to make meaningful statements about all of these parameters simultaneously. Below, we shall attempt to provide two-dimensional cross-sections of this 13-dimensional space and give graphical illustrations of the variability of the design variances, and hence the domains of optimality of each design, as we vary two parameters at a time. We provide the graphs of variances of the designs involved (typically, the cluster-panel design with dotted lines, the observation-panel design with dashed lines, and the independent design with dash-dotted lines. For most plots, the independent design is not affected by the variations of the parameters that make up the axis of the plots, and hence omitted). We also show the relative efficiency of different designs, marking the domains of the parameter space in yellow/light gray if the independent design is the most efficient one; in green/medium gray if the cluster-panel design is the most efficient one; and in purple/dark gray if the observation-panel design is the most efficient one (R code used to produce graphs is available at <http://web.missouri.edu/~kolenikovs/SMJ2011/>).

Figure 1 shows how the design variances, and hence the most efficient design, vary with the panel costs of the PSU and SSU, c_{12}^I and c_{12}^{II} . Obviously, these variations do not affect the variance of the independent design, which serves as a benchmark. Also, the variations in c_{12}^{II} do not affect the performance of the cluster-panel design, which corresponds to the dotted vertical iso-variance lines on the left panel. The dashed downward sloping lines are the iso-variance lines for the observation-panel design. Note that the lower left corner of the graph corresponds to the free lunch situation in which the second wave of data collection does not cost anything: the panel costs are equal to the single period cost, $c_{12}^I = c_1^I$, $c_{12}^{II} = c_1^{II}$. When the costs of the panel data collection are prohibitively high (the upper right corner of the graph), the independent design is the most efficient one. The point where all three designs have the same variances is $c_{12}^I = 22$, $c_{12}^{II} = 3.05$; i.e., the cost of the second interview is 2.05 higher than the cost of the first interview, and the cluster-level costs in the second wave are 20% higher than in the first wave. Still, a positive autocorrelation justifies the reduction in the sample size of the observation-panel design as compared to the independent design. If the cluster level panel cost is lower and the second interview cost is higher, the cluster-panel design is the most efficient. For inexpensive second interviews, the most efficient design is the observation-panel design. The latter domain includes our baseline case with $c_{12}^I = 18$ and $c_{12}^{II} = 3$.

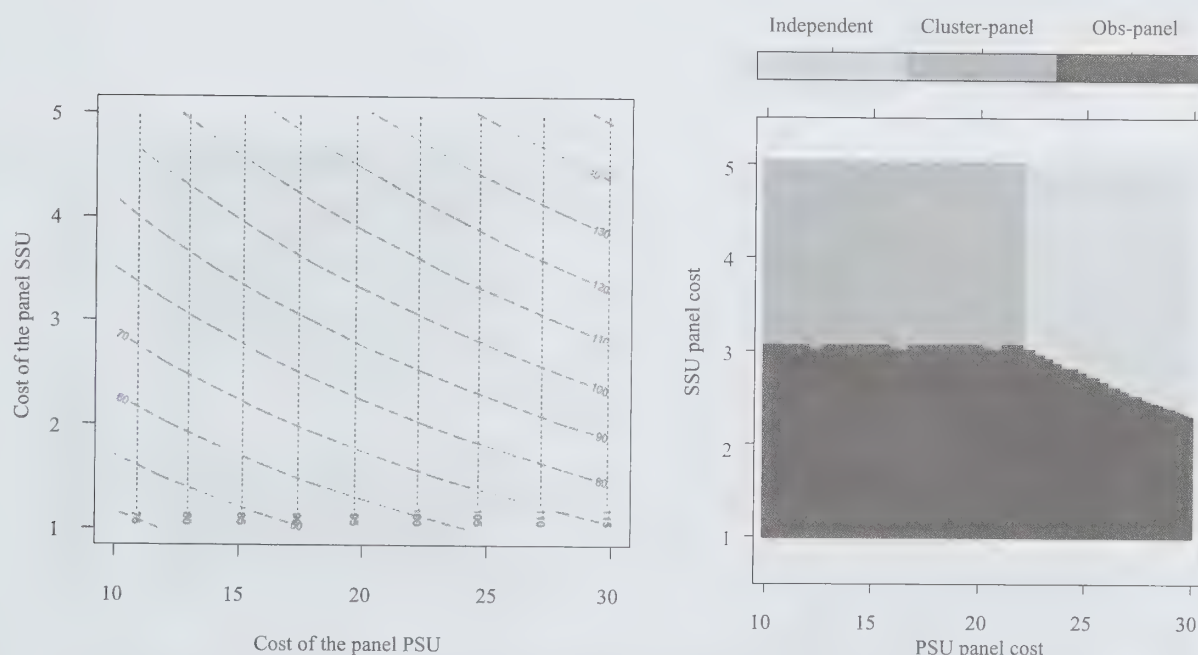


Figure 1 Design variances as functions of the data collection costs c_{12}^I, c_{12}^{II} . Left: contour lines of $V_{e,c}[d]$ (dotted) and $V_{e,o}[d]$ (long dashed); $V_{e,i} = 99.86$; right: domains of optimality of the three designs

Figure 2 shows the changes in design variances associated with the changes in the autocorrelations ρ^I, ρ^{II} . The independent design variance is unaffected by these variations, and the cluster-panel design is unaffected by variations in ρ^{II} . The observation-panel design is more efficient for higher SSU autocorrelation, $\rho^{II} > 0.34$. Otherwise, the cluster-panel design provides lower variance.

Figure 3 investigates the impact of the cluster-level cost and autocorrelation on the choice of the design. The combinations of expensive second wave of data collection and low PSU autocorrelation in the upper left corner of the plot makes the independent design the most appealing one. Otherwise, the observation-panel design is the best one to use. Note that the contour lines for the cluster-panel and observation-panel designs are very close to one another, and differences in variances between the two designs are less than 2% in the whole parameter space of this plot.

Figure 4 investigates the impact of the observation-level cost and autocorrelation on the choice of the design. Neither the independent design nor the cluster-panel design variances are affected by variation of the parameters shown on this plot. The independent design variance is 99.86, while the cluster-panel design variance is 91.37, so the observation-panel design is compared to the latter only. High autocorrelations ($\rho^{II} \geq 0.6$) can justify very high cost of the second interview (up to fourfold compared to the first interview), but in the upper left corner of the plot corresponding to the low autocorrelations and high panel cost, the cluster-panel design performs better.

Figure 5 relates the design variances to the cluster-level costs of the survey. The horizontal axis is the cost in the first period, c_1^I , and the vertical axis is the additional cost of in the second period when the data are collected in a panel mode, $c_{12}^I - c_1^I$. The vertical axis is ignored for the independent design, as this parameter does not appear in the independent design. Also, by virtue of (4.1), $c_1^I = c_2^I$. The observation-panel design is uniformly better than the cluster-panel design for all parameter combinations on this graph, although the difference in variances does not exceed 2%. In the upper left corner, the additional cost of the panel mode of data collection is prohibitively high, and the independent design offers better performance.

Figure 6 shows the dependence of the most efficient design on the total budget of the survey and the cost of panel mode of data collection at the cluster level. For $C_0 > 10,000$, the observation-panel design performs better if $c_{12}^I < 22.7$, i.e., if the additional cost of the panel mode of data collection at the cluster level does not exceed 127% of the initial cluster-level cost in the first wave. Interestingly, for some isolated parameter configurations in small surveys, the cluster-panel design can perform better than the observation-panel design that dominates the rest of the plot. The difference in design variances between the cluster-panel and observation-panel designs is less than 4% across all parameter combinations on this graph.

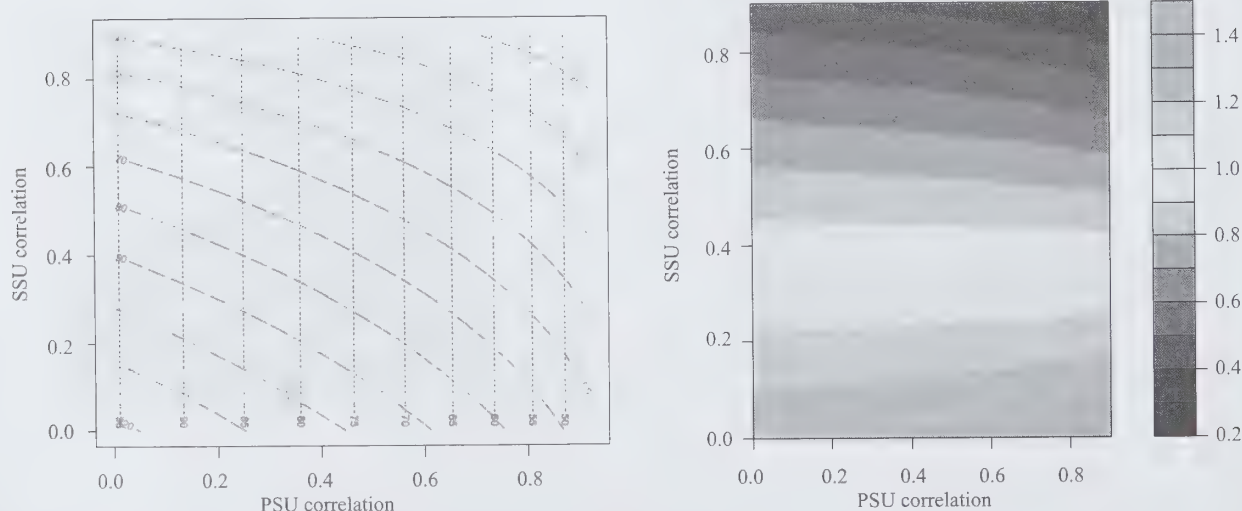


Figure 2 Design variances as functions of the population correlations ρ^I, ρ^{II} . Left: contour lines of $V_{e,c}[d]$ (dotted) and $V_{e,o}[d]$ (long dashed); $V_{e,i} = 99.86$; right: ratio $V_{e,o}[d] / V_{e,c}[d]$

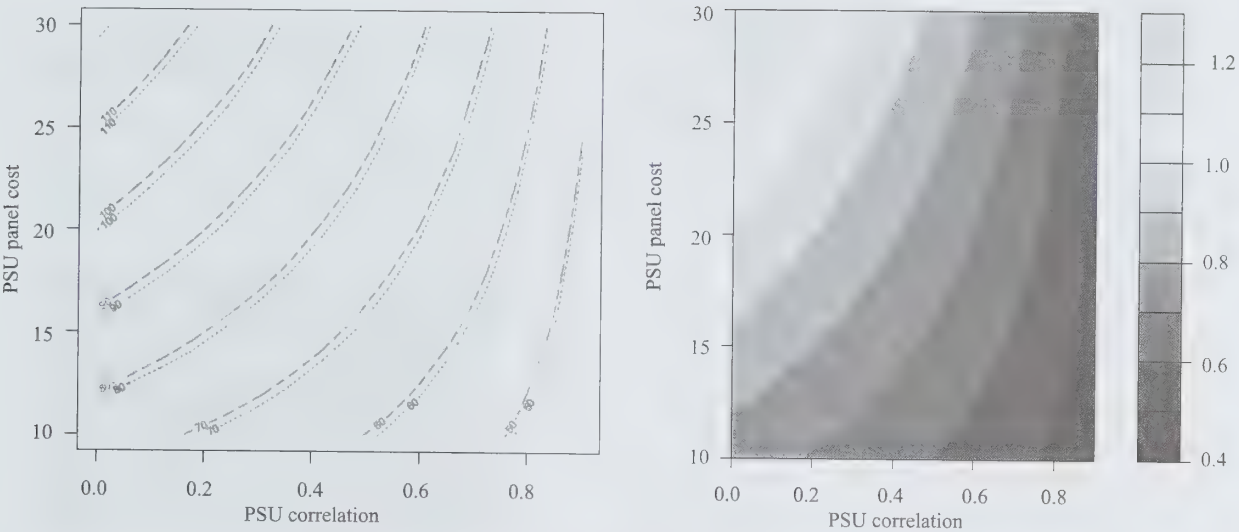


Figure 3 Design variances as functions of the cluster-level autocorrelation ρ^I and cost c_{12}^I . Left: contour lines of $V_{e,c}[d]$ (dotted) and $V_{e,o}[d]$ (long dashed); $V_{e,l} = 99.86$; right: ratio $V_{e,o}[d] / V_{e,l}[d]$

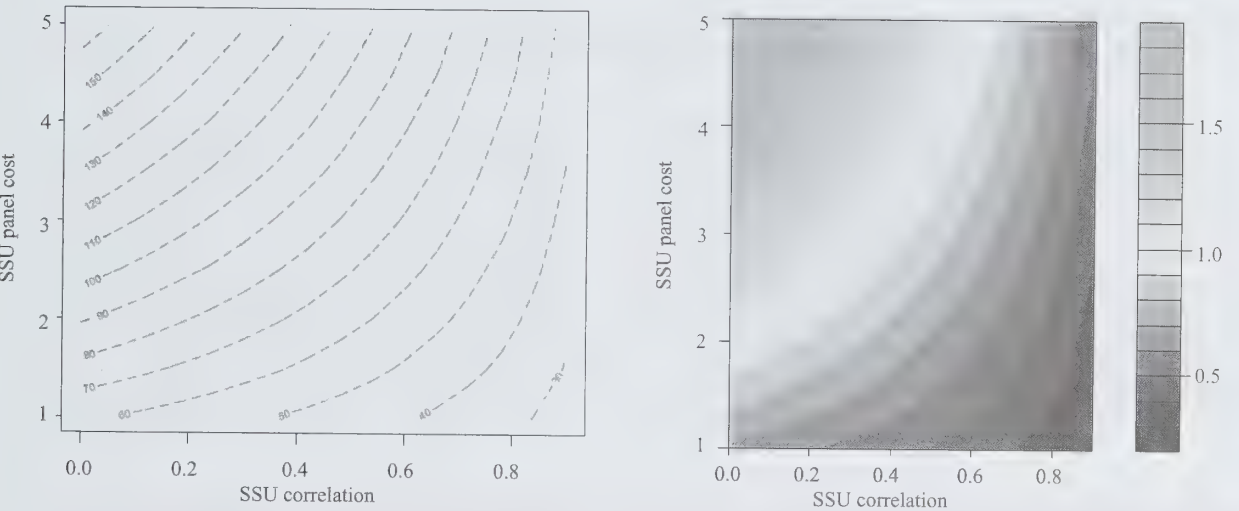


Figure 4 Design variances as functions of the observation-level autocorrelation ρ^{II} and cost c_{12}^{II} . Left: contour lines of $V_{e,o}[d]$ (long dashed); $V_{e,l} = 99.86$; $V_{e,c}[d] = 91.37$; right: ratio $V_{e,o}[d] / V_{e,c}[d]$

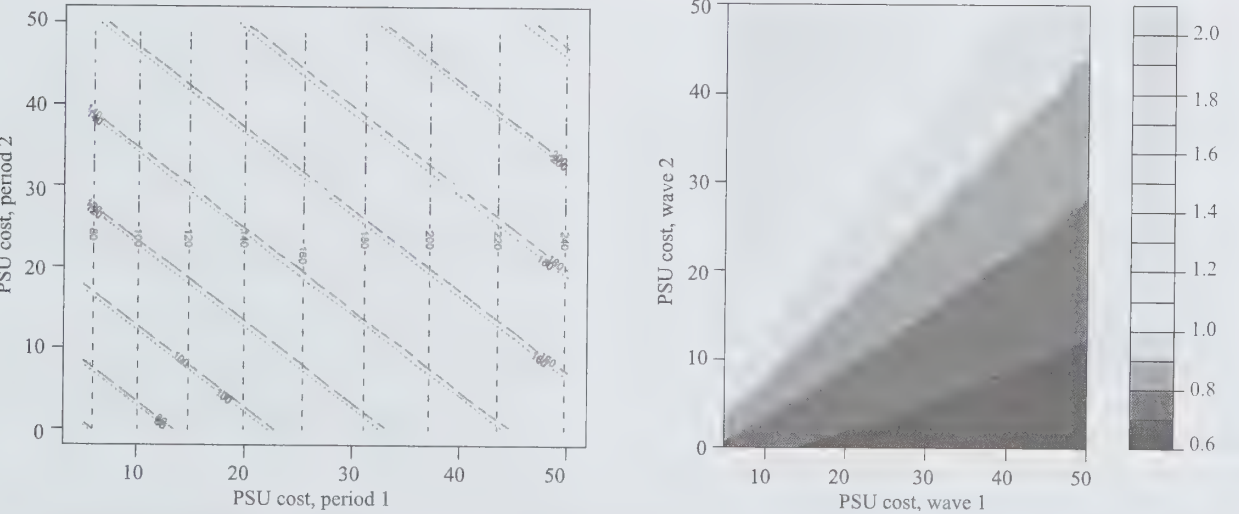


Figure 5 Design variances as functions of the cluster level costs in the first wave, c_1^I , and in the second wave, $c_{12}^I - c_1^I$. Left: contour lines of $V_{e,c}[d]$ (dotted), $V_{e,o}[d]$ (long dashed) and $V_{e,l}[d]$ (dash-dotted); right: ratio $V_{e,o}[d] / V_{e,l}[d]$

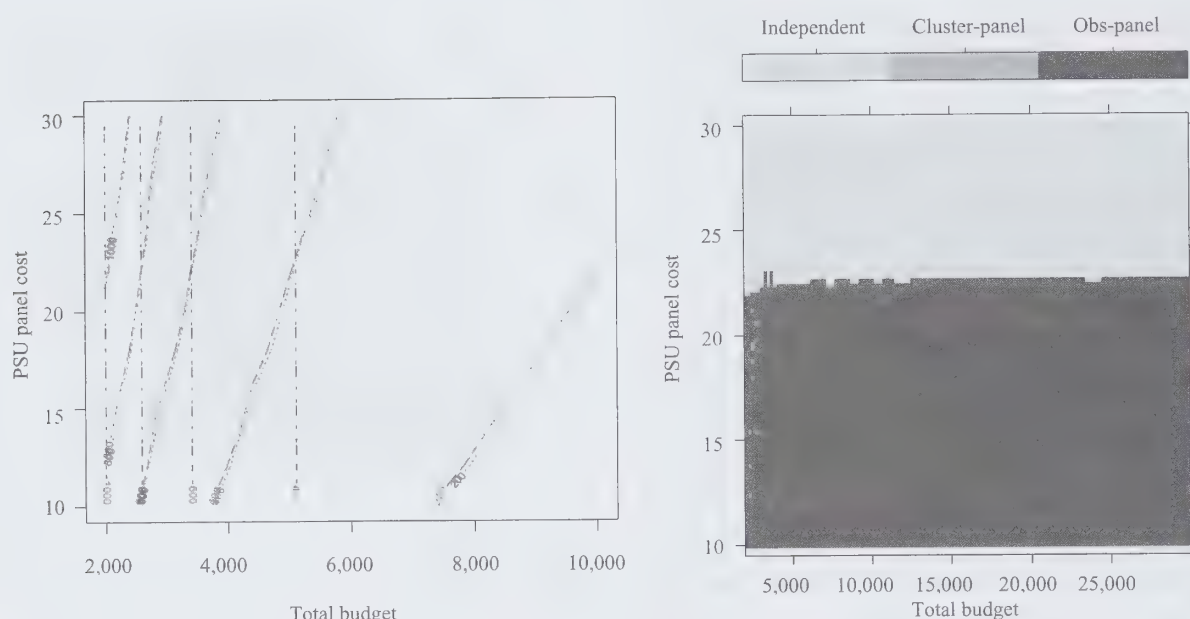


Figure 6 Design variances as functions of the total budget C_0 and the PSU panel cost c_{12}^H . Left: contour lines of $V_{e,c}[d]$ (dotted), $V_{e,o}[d]$ (long dashed) and $V_{e,i}[d]$ (dash-dotted); right: domains of optimality of the three designs

Overall, this numeric illustration shows that depending on the parameters of the population and costs of data collection, each of the three designs can be the most efficient one. Low correlations and high costs in the second wave tend to favor the independent design. Given that the initial six population parameters and five cost parameters may not be representative of many repeated surveys, a sensitivity analysis like the one performed here may be needed for any particular survey a statistician needs to design.

5. Survey design with multiple criteria

So far, our analysis was confined to estimation of the difference between the means in two waves of data collection of a single variable. Most large scale surveys are collected to study several characteristics, and to many users, the contemporaneous estimates are also of interest. To accommodate accuracy requirements associated with these different variables and different estimates, the survey designer must have several variances in mind when choosing the design to be implemented. This is a multicriterial optimization problem, and no single design will work best for all possible estimation problems. In the current context, the observation-panel design may give good estimates of the change when both PSU and SSU autocorrelations are high, but it may result in a small sample size if both PSUs and SSUs are expensive to follow up. Greater precision of the estimates for any single period could be obtained by switching to the cluster-panel or even independent designs.

Comparing different designs in this situation is possible with the standard microeconomic argument of utility maximization under budget constraints (Mas-Colell, Whinston and Green 1995). In the survey design context, the utility of the survey designer increases with the precision of the survey estimates, or equivalently decreases with survey variances. A simple functional form is given by Cobb-Douglas utility function:

$$U(\text{design}) = V_{\text{design}}^{-\alpha_1}[\bar{y}_1] V_{\text{design}}^{-\alpha_2}[\bar{y}_2] V_{\text{design}}^{-\alpha_3}[d]. \quad (5.1)$$

Here, α_1 , α_2 and α_3 are positive constants describing the relative weights of the three design variances in decision-making process. Variances $V[\bar{y}_1]$ and $V[\bar{y}_2]$ in (5.2) are the variances of the means in cluster surveys given by (2.8). The variance of the difference estimator is (2.10), (2.12) or (2.14), depending on the design. The survey designer problem is then to maximize (5.1) subject to design-specific budget constraints (3.1), (3.4) or (3.6). Maximization is performed over the design parameters (mode of data collection, number of clusters in each time period, number of observations in each time period), given the characteristics of population (variances and autocorrelations) and the data collection process (costs).

Let us assume that the precision of each of the three estimates \bar{y}_1 , \bar{y}_2 and d is equally important to the decision maker, so $\alpha_1 = \alpha_2 = \alpha_3$. To have an objective function that is measured in the variance units and is on the same scale as variances, it will be convenient to define a multicriterial variance

$$V_{\text{design}} = (V_{\text{design}}[\bar{y}_{1..}] V_{\text{design}}[\bar{y}_{2..}] V_{\text{design}}[d])^{1/3}, \quad (5.2)$$

and express the optimization problem as minimization of this expression.

Analytic characterization of the design that optimizes (5.2) becomes quite cumbersome. Instead, we utilize a numeric illustration of the previous section to demonstrate how accounting for other design objectives affects the choice of the design. We should expect that for the designs with more expensive follow-ups ($c_{12}^I \geq c_1^I + c_2^I$, $c_{12}^{II} \geq c_1^{II} + c_2^{II}$), the simpler designs would be selected more often: the cluster-panel design may be preferred to the observation-panel design, and the independent design may be preferred to the cluster-panel design. For the baseline settings (4.1), we have

$$V_{e,l}[\bar{y}] = 49.93, V_{e,c}[\bar{y}] = 47.68, V_{e,o}[\bar{y}] = 61.69,$$

$$V_{e,l} = 62.91, \quad V_{e,c} = 59.23, \quad V_{e,o} = 70.02,$$

where the time indices of $y_{t..}$ are omitted. The observation-panel design is rather inefficient in estimating the period-specific means as this design samples fewer units. Instead, the cluster-panel design is the most efficient one, closely followed by the independent design.

Figures 7-12 parallel Figure 1-6, respectively. Since the best design in terms of V is now the cluster-panel design, most of these plots show the preference toward this design. Figure 7 shows that when the variances of the contemporaneous means are taken into account, the simpler independent and cluster-panel designs are preferred for a greater fraction of parameter settings, and occupy a larger portion of

the plot than in Figure 1. The point where the three designs are equivalent is $c_{12}^I = 20.6$, $c_{12}^{II} = 2.27$, closer to the origin than in Figure 1, in which only the variance of the difference was taken into account.

Figure 8 shows that the observation-panel design is only justified when both autocorrelations are higher than 0.6 (for the given values of population variances and costs). Recall that in Figure 2, the observation-panel design was preferred whenever $\rho^{II} > 0.34$, with little dependence on ρ^I .

Figure 9 shows how the PSU-level correlations and costs affect the choice of the design. The observation-panel design is less efficient than the cluster-panel design for all combinations of parameters in this plot. Hence, the choice of the design is between the independent and the cluster-panel designs. Naturally, if the data collection in the panel mode is expensive, the independent design is preferred to the cluster-panel design. Interestingly, the preference towards a particular design is not monotone in ρ_{12}^I . With values $\rho_{12}^I > 0.7$, the $V[d]$ component in (5.2) produces designs with so few clusters that $V[\bar{y}]$ suffers notably enough to hurt the whole objective function. At that value of panel autocorrelation, the maximum panel cost at which the cluster-panel design is still the most efficient one is $c_{12}^I = 24.4$, *i.e.*, the cluster-level cost in the second wave is 44% higher than in the first wave.

Figure 10 shows that the higher autocorrelation of the SSU measurements may justify modest extra cost associated with data collection. The highest cost for which the observation-panel design is still the most efficient one is $c_{12}^{II} = 2.75$ with $\rho^{II} = 0.78$; *i.e.*, the cost of the second interview can be 75% more than the cost of the first interview.

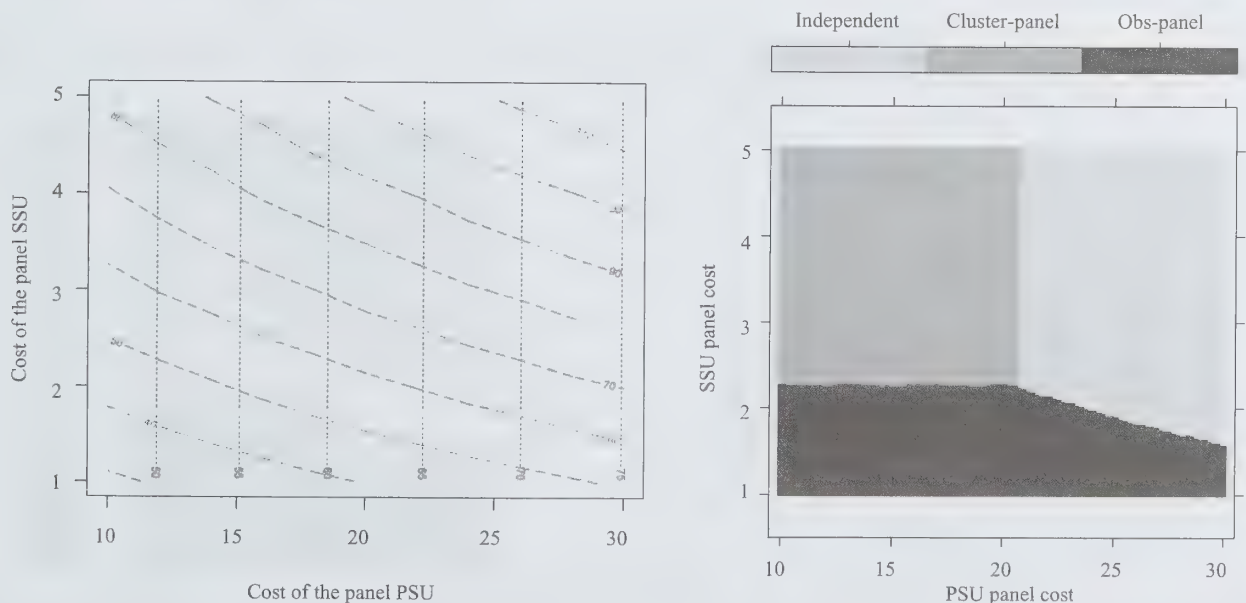


Figure 7 Design variances as functions of the data collection costs c_{12}^I, c_{12}^{II} . Left: contour lines of $V_{e,c}$ (dotted) and $V_{e,o}$ (long dashed); $V_{e,l} = 62.91$; right: domains of optimality of the three designs

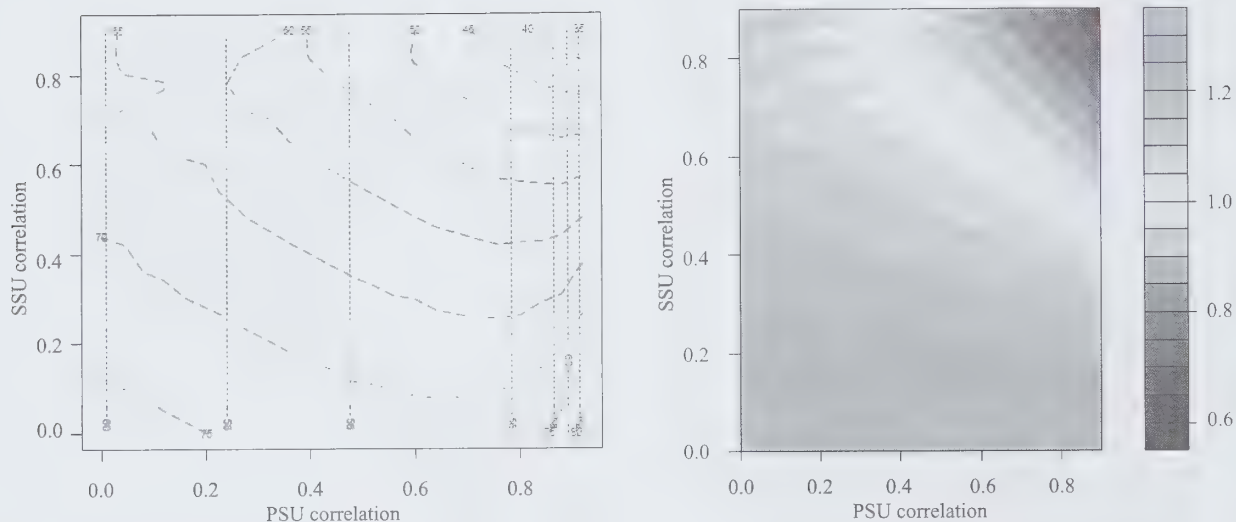


Figure 8 Design variances as functions of the autocorrelations ρ^I, ρ^{II} . Left: contour lines of $V_{e,c}$ (dotted) and $V_{e,o}$ (long dashed); $V_{e,l} = 62.91$; right: ratio $V_{e,o}/V_{e,c}$

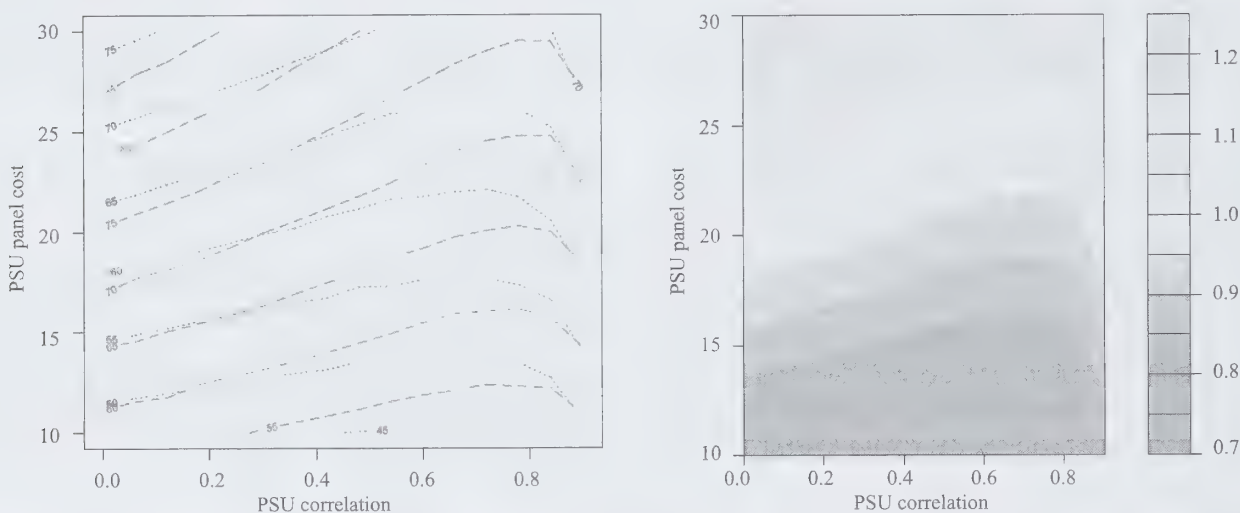


Figure 9 Design variances as functions of the cluster-level autocorrelation ρ^I and cost c_{12}^I . Left: contour lines of $V_{e,c}$ (dotted) and $V_{e,o}$ (long dashed); $V_{e,l} = 62.91$; right: ratio $V_{e,c}/V_{e,l}$

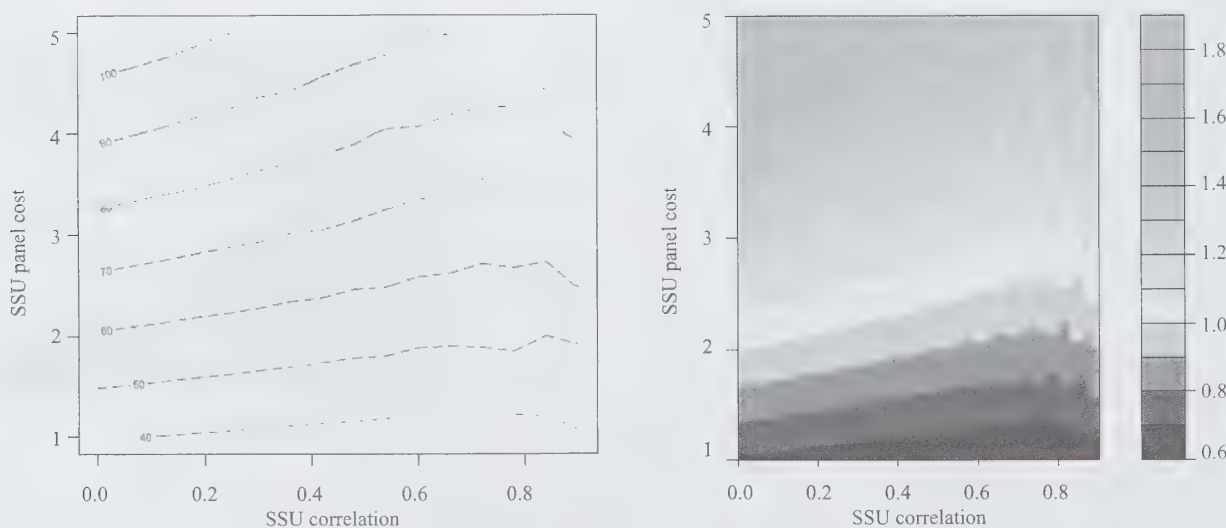


Figure 10 Design variances as functions of the observation-level autocorrelation ρ^{II} and cost c_{12}^{II} . Left: contour lines of $V_{e,o}$ (long dashed); $V_{e,l} = 62.91$; $V_{e,c} = 59.23$; right: ratio $V_{e,o}/V_{e,c}$

Figure 11 parallels Figure 5. The left panel shows that the observation-panel design is less efficient than the cluster-panel design. The right panel shows that if the cluster-level cost of the second wave exceeds the cluster-level cost of the first wave by more than 15 units, the independent design delivers better efficiency than the cluster-panel design.

Finally, Figure 12 shows the variances as functions of the total survey budget and the cost of the panel mode of data collection. There is very little dependence on C_0 in the plot, and the independent design is preferred if the panel mode is too expensive, namely, when the cluster-level cost in the second cost exceeds 107% of that in the first wave.

As it was conjectured in the beginning of this section, incorporation of the variances of the contemporaneous means into the design optimization objective function shifted the preferences of the survey designer towards simpler designs that can sample a greater number of the ultimate observation units. The observation-panel design now only makes sense when both the PSU and SSU autocorrelations are high, and the panel costs are reasonably low. Moreover, the cluster-panel design is generally justified only if there is an economy in cluster-level cost in the second wave of the survey.

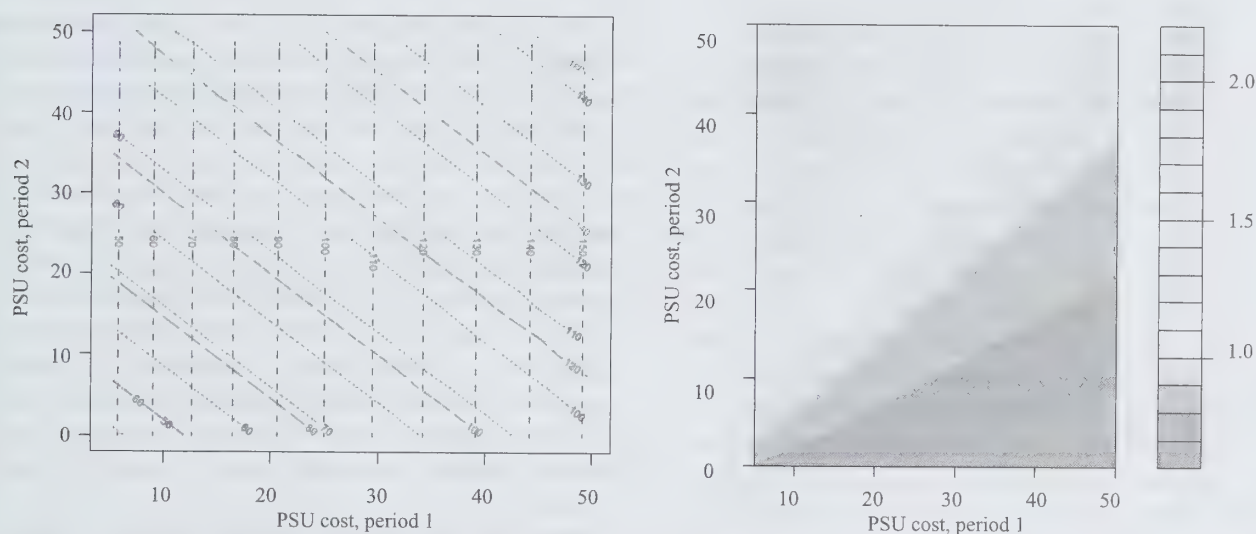


Figure 11 Design variances as functions of the data collection costs c_1^I, c_{12}^I . Left: contour lines of $V_{e,c}$ (dotted), $V_{e,o}$ (long dashed) and $V_{e,u}$ (dash-dotted); right: ratio $V_{e,c}/V_{e,u}$

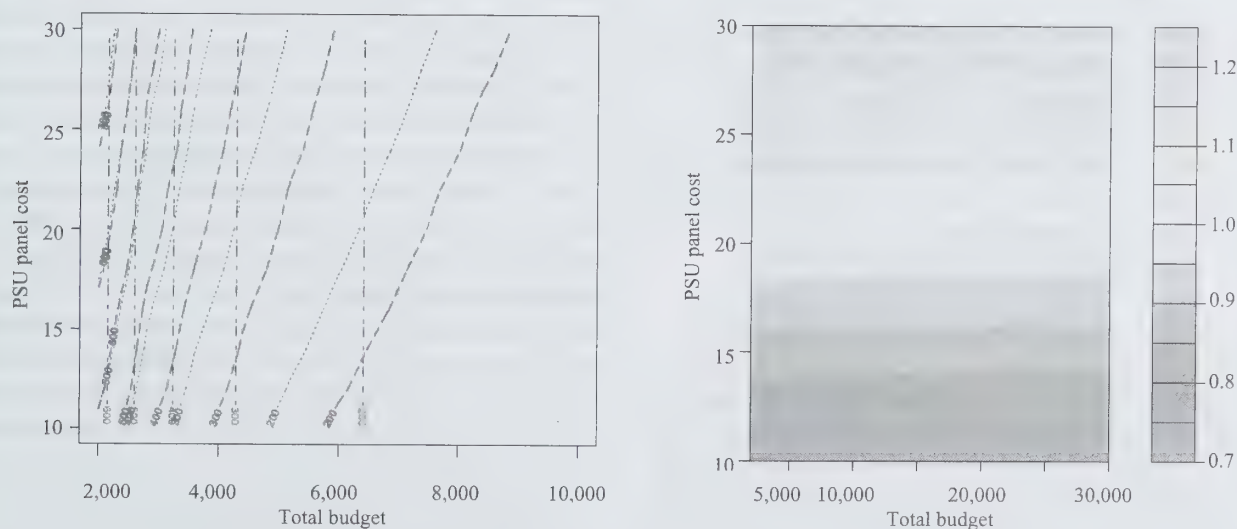


Figure 12 Design variances as functions of the total budget C_0 and the PSU panel cost c_{12}^{II} . Left: contour lines of $V_{e,c}$ (dotted), $V_{e,o}$ (long dashed) and $V_{e,u}$ (dash-dotted); right: domains of optimality of the three designs

6. Extensions to multiple waves

If the survey to be designed will have more than two waves of data collection, the survey designer may be able to extend the framework of the utility maximization problem (5.1), with the following considerations in mind.

1. A greater number of targets of inference. Possible variances that the survey designer may need to take into account can now include: contemporaneous variances $V[\bar{y}_1], V[\bar{y}_2], \dots, V[\bar{y}_T]$; consecutive differences $V[\bar{y}_2 - \bar{y}_1], \dots, V[\bar{y}_T - \bar{y}_{T-1}]$ or composite/GLS estimators of the change between two adjacent periods of time; other contrasts $V[\sum_t c_t \bar{y}_t], \sum c_t = 0$; variance of the linear growth rates from regression of \bar{y}_t on t , estimated by OLS or GLS; *etc.*
2. A possibility of discounting. In economics, it is customary to specify the budget constraints that look into the future in the form of $\sum_t x_t \delta^t$ where x_t is the amount spent in time t , and $\delta < 1$ is the discount factor associated with interest rates. Discounting may also be relevant for the utility function, and design variances farther in the future may have lower weights in the optimization problem.
3. Unknown functional forms of the time-series processes associated with the variable of interest. The survey designer needs to have a good idea about the covariance structure of the time series of both individual observations and cluster means. It is likely that the results will be sensitive to the choice of the particular model. In the current analysis, the issue is ameliorated, as it suffices to have a single correlation parameter for each level. The survey designer may have to introduce more parameters into the model, and correspondingly study sensitivity of the design choice with respect to these parameters.

The complexity of the problem, as outlined above, can grow out of control very quickly. We thus abstain from a more detailed treatment of it in this paper.

7. Discussion

This paper has analyzed different options for implementation of repeated cluster surveys. We have provided analytical expression for design variances of the simple difference estimator for three popular designs (the independent, the cluster-panel and the observation-panel designs). We have also derived the optimal sample sizes for estimation of the difference between two waves of data collection.

The sample designer who knows that the characteristic of interest is going to have some degree of persistence over time will likely choose one of the panel designs, provided that the costs of re-visiting the clusters and/or observation units are not prohibitively high. Analytical comparison is possible between the independent and the observation-panel designs, and is given by Proposition 7. It is worth noting that the design variance of the difference is $O(C_0^{-1})$ for both the independent design and the observation-panel design, and is $O(C_0^{-1/2})$ for the cluster-panel design, where C_0 is the total budget of the survey. Hence the cluster-panel design is only viable for smaller surveys, while the large scale surveys will likely have either the independent or the observation-panel format.

The cost structure considered in Section 3 is rather simplistic. For instance, the second stage costs in the second time period may differ across individuals sampled from the new or from the reused clusters. Also, the costs may depend on the cluster size M_i , as it may take more time and resources to obtain maps and collect cluster level data for bigger clusters. Our original motivation was to consider situations in which the SSU panel cost is higher than twice the cost of individual interviews. However, as suggested by one of the referees, this cost may be lower if the follow-up interviews are performed in cheaper mode, such as a phone interview or a self-administered mail survey instead of a personal interview. If this is the case, the observation-panel design is apparently the most cost-efficient of the three designs.

The population structure is also an oversimplification. The clusters are assumed to be of balanced unchanging sizes. No units leave the population, and no new units appear. These assumptions are quite restrictive for many practical situations. If the population changes between two waves of data collection, the sample designer would want to include new clusters at the second wave, using the algorithms of Ernst (1999). The new clusters are placed into a separate stratum, and a clustered sample is taken from that stratum. In NHIS, this is implemented by "permit" frame. Also, the dynamic measurement effects such as conditioning and time in sample lead to rotation bias, so it might be beneficial to provide at least some rotation of the PSUs. For DHS studies, in particular, the first argument (coverage) is likely to be more important than the second one (time in sample) due to a substantial time between the waves of the survey (about 5 years). Arguably, both non-response and loss of coverage can be added to the current framework as sources of bias, leading to optimization of the mean squared total survey error rather than the design variance. Convincing models of such biases may be difficult to formulate, however.

Another issue that would arise with clusters of different sizes is that of the greater range of applicable designs. In this paper, we assumed SRSWOR at both stages. Other designs, such as sampling with probability proportional to size (PPS), can be used instead. For designs other than SRS, the Horvitz-Thompson estimator and its variance (Sämdal, Swensson and Wretman 1992, Thompson 1997) would need to be used. The analytical derivations become unwieldy, although practical numerical demonstrations similar to our Sections 4 and 5 can still be implemented. If cluster sizes change over time, obtaining the optimal design becomes a moving target, and designs optimal for the “old” measures of size will lose their efficiency with the “new” measures of size.

In earlier drafts of this paper, we analyzed intermediate designs where a non-trivial fraction of the units are retained, and other units are sampled independently. The problem can then be viewed as variance minimization subject to inequality constraints on the degree of the overlap $0 \leq \pi^I \leq 1$, $0 \leq \pi^{II} \leq 1$. The general theory of non-linear constrained optimization ensures that as long as the variance of the population mean change D is monotone in π^I and π^{II} , the optimum will be achieved in one of the vertices of the parameter space. This justifies our interest in the three designs considered in the paper. They correspond to the vertices of the parameter space: $(0, 0)$, $(1, 0)$ and $(1, 1)$ for the independent, cluster-panel and observation-panel designs, respectively. The point $(0, 1)$ corresponds to an impossible design with complete overlap of the individual units with no overlap of the clusters. Cumbersome derivations show that it is possible to satisfy the first order conditions in some intermediate cases, too, but they correspond to local maxima of the variance. While these results may also be of interest (in the sense of providing an upper bound on the design variances), we did not consider them in the paper. In the more complicated cases of the multicriterial optimization of Section 5, monotonicity does not necessarily hold, and other designs beside the three extreme cases considered in the paper may lead to the optimal values of the objective function (5.2).

Conditions of equal variances (2.9) can be relaxed at the price of producing substantially more complicated expressions. If the sample sizes are fixed between the two occasions, then the following changes will be necessary in all relevant formulas. In the expressions that do not involve autocorrelations,

$$2S_h^2 \mapsto S_{1h}^2 + S_{2h}^2, \quad 2S_w^2 \mapsto \bar{S}_{1w}^2 + \bar{S}_{2w}^2, \quad (7.1)$$

while in the expressions that do involve autocorrelations,

$$\begin{aligned} 2(1 - \rho^I)S_b^2 &\mapsto S_{1b}^2 + S_{2b}^2 - 2\rho^I S_{1b}S_{2b}, \\ 2S_w^2(1 - \rho^{II}) &\mapsto \bar{S}_{1w}^2 + \bar{S}_{2w}^2 - 2\rho^I \bar{S}_{1w}\bar{S}_{2w}. \end{aligned} \quad (7.2)$$

Qualitatively, the results will be the same.

The multicriterial framework of Section 5 allows for different importance weights to be given to different variances of interest. Relatively larger values of α_1 , α_2 correspond to the greater importance of the contemporaneous means, while larger values of α_3 correspond to the greater importance of the change estimate. The original problem of optimizing the design for $V[d]$ can be considered within the context of (5.1) by setting $\alpha_1 = \alpha_2 = 0$, $\alpha_3 = 1$. This framework can also be expanded to include designs aimed at measuring several variables. An additional challenge of such a setup is that the autocorrelations may differ across different variables. Some individual characteristics are constant over time (race, gender); others change slowly (housing, expenditure, political preferences), yet others may change faster (income or behavior).

This paper dealt with three designs and a specific estimator of change: the difference in the two estimates of the mean in two periods of time. Other options for either designs or estimators are also available. For instance, in rotation designs, a fraction of the first wave units is retained, and some new units are recruited. For such designs, composite estimation (Hansen *et al.* 1953, Patterson 1950, Rao and Graham 1964, Wolter 2007) that weighs differently the contributions of the independent units (those retired from the sample after the first wave, and those newly recruited for the second wave) and the contributions of the panel units (used in both waves) would result in more efficient estimates. Generally, motivation for such designs comes from non-sampling considerations, such as decrease of the response burden and deterioration of the sample representativeness of population due to the population change. These considerations can be accounted for in either the cost model (e.g., a greater number of callbacks required to convince a unit to respond), or the total survey error model (by introducing the non-response or undercoverage bias, and considering mean squared error rather than the design variance of an estimate).

Acknowledgements

The authors are grateful to Chris Skinner and John Eltinge for helpful discussions, to William Kalsbeek for suggestions at the early stages of the paper, and to the associate editor and two referees for their comments. Nash Herndon and Oksana Loginova provided editorial improvements. Partial financial support was provided by U.S. Agency for International Development through the MEASURE Evaluation project of Carolina Population Center, University of North Carolina at Chapel Hill, under the terms of Cooperative Agreement GPO-A-00-03-00003-00. The authors are also grateful to the participants of the Joint Statistical Meetings

(2005) and the XXIII International Methodology Symposium of Statistics Canada (2007) for helpful comments.

Appendix

Expectations, variances and covariances in the proofs below are with respect to the corresponding designs. The first stage of selection will be denoted with a superscript I. The second stage of selection will be denoted with a superscript II.

Proof of Proposition 2. Let us denote the sample of the PSUs by S^I , the sample of SSUs in the first period by S_{i1}^{II} , and the sample of SSUs in the second period by S_{i2}^{II} . Then

$$d = \bar{y}_{2..} - \bar{y}_{1..} = \frac{1}{mn} \sum_{i \in S^I} \left(\sum_{j \in S_{i2}^{II}} y_{2ij} - \sum_{j \in S_{i1}^{II}} y_{1ij} \right).$$

Denoting the expectations with respect to the first stage as E_I , and those with respect to the second stage as E_{II} , we have the design variance of d equal to

$$\begin{aligned} V[d] &= E_I V_{II}[d | S^I] + V_I E_{II}[d | S^I] \\ &= \frac{1}{m^2 n^2} E_I \left\{ \sum_{i \in S^I} V_{II} \left[\sum_{j \in S_{i2}^{II}} y_{2ij} - \sum_{j \in S_{i1}^{II}} y_{1ij} \right] \right\} \\ &\quad + \frac{1}{m^2 n^2} V_I \left\{ \sum_{i \in S^I} E_{II} \left[\sum_{j \in S_{i2}^{II}} y_{2ij} - \sum_{j \in S_{i1}^{II}} y_{1ij} \right] \right\} \\ &= \frac{1}{m^2 n^2} E_I \left\{ \sum_{i \in S^I} V_{II} \left[\sum_{j \in S_{i2}^{II}} y_{2ij} \right] + V_{II} \left[\sum_{j \in S_{i1}^{II}} y_{1ij} \right] \right\} \\ &\quad + \frac{1}{m^2 n^2} V_I \left[\sum_{i \in S^I} m \bar{Y}_{2i.} - m \bar{Y}_{1i.} \right] \\ &= \frac{1}{m^2 n^2} E_I \left[\sum_{i \in S^I} \left(1 - \frac{m}{M} \right) m S_{2wi}^2 + \left(1 - \frac{m}{M} \right) m S_{1wi}^2 \right] \\ &\quad + \frac{1}{n^2} \left(1 - \frac{n}{N} \right) n (S_{1b}^2 + S_{2b}^2 - 2\rho^I S_{1b} S_{2b}) \\ &\quad + \frac{1}{m^2 n^2} nm \left(1 - \frac{m}{M} \right) m (S_{2w}^2 + S_{1w}^2) \\ &\quad + \frac{1}{n} \left(1 - \frac{n}{N} \right) (S_{1b}^2 + S_{2b}^2 - 2\rho^I S_{1b} S_{2b}) \\ &= \left(1 - \frac{n}{N} \right) \frac{2S_b^2(1 - \rho^I)}{n} + \left(1 - \frac{m}{M} \right) \frac{2S_w^2}{mn}, \end{aligned}$$

where the last equality assumes symmetric conditions (2.9).

Proof of Proposition 3. Let us denote the sample of the PSUs by S^I , and the sample of SSUs, by S_i^{II} . Then

$$d = \bar{y}_{2..} - \bar{y}_{1..} = \frac{1}{mn} \sum_{i \in S^I} \sum_{j \in S_i^{II}} (y_{2ij} - y_{1ij}).$$

Denoting the expectations with respect to the first stage as E_I , and those with respect to the second stage as E_{II} , we have the design variance of d equal to

$$\begin{aligned} V[d] &= E_I V_{II}[d | S^I] + V_I E_{II}[d | S^I] \\ &= \frac{1}{m^2 n^2} E_I \left[\sum_{i \in S^I} V_{II} \sum_{j \in S_i^{II}} (y_{2ij} - y_{1ij}) \right] \\ &\quad + \frac{1}{m^2 n^2} V_I \left[\sum_{i \in S^I} E_{II} \sum_{j \in S_i^{II}} (y_{2ij} - y_{1ij}) \right] \\ &= \frac{1}{m^2 n^2} E_I m \left[\sum_{i \in S^I} \left(1 - \frac{m}{M} \right) (S_{2wi}^2 + S_{1wi}^2 - 2S_{2wi} S_{1wi} \rho^{II}) \right] \\ &\quad + \frac{1}{m^2 n^2} V_I \left[\sum_{i \in S^I} m (\bar{Y}_{2i.} - \bar{Y}_{1i.}) \right] \\ &= \frac{1}{mn^2} n \left(1 - \frac{m}{M} \right) (\bar{S}_{1w}^2 + \bar{S}_{2w}^2 - 2\rho^{II} \bar{S}_{1w} \bar{S}_{2w}) \\ &\quad + \frac{1}{n^2} n \left(1 - \frac{n}{N} \right) (S_{1b}^2 + S_{2b}^2 - 2\rho^I S_{1b} S_{2b}) \\ &= \left(1 - \frac{n}{N} \right) \frac{S_{1b}^2 + S_{2b}^2 - 2\rho^I S_{1b} S_{2b}}{n} \\ &\quad + \left(1 - \frac{m}{M} \right) \frac{\bar{S}_{1w}^2 + \bar{S}_{2w}^2 - 2\rho^{II} \bar{S}_{1w} \bar{S}_{2w}}{mn} \\ &= 2 \left(1 - \frac{n}{N} \right) \frac{S_b^2(1 - \rho^I)}{n} \\ &\quad + 2 \left(1 - \frac{m}{M} \right) \frac{S_w^2(1 - \rho^{II})}{mn}, \end{aligned}$$

with the last equality holding under the symmetry conditions.

Proof of Proposition 4. The Lagrangian function of minimizing (2.11) subject to constraint (3.1) is

$$\begin{aligned} L(n_1, m_1, n_2, m_2, \lambda) &= \\ &\quad \left(1 - \frac{n_1}{N} \right) \frac{S_b^2}{n_1} + \left(1 - \frac{n_2}{N} \right) \frac{S_b^2}{n_2} \\ &\quad + \left(1 - \frac{m_1}{M} \right) \frac{\bar{S}_w^2}{n_1 m_1} + \left(1 - \frac{m_2}{M} \right) \frac{\bar{S}_w^2}{n_2 m_2} \\ &\quad - \lambda (c_1^I n_1 + c_1^{II} n_1 m_1 + c_2^I n_2 + c_2^{II} n_2 m_2 - C_0). \end{aligned}$$

Working through the first order conditions of this Lagrangian function leads to

$$\begin{aligned}
 -\lambda &= \frac{m_1 S_b^2 + \left(1 - \frac{m_1}{M}\right) \bar{S}_w^2}{n_1^2 m_1 (c_1^I + c_1^{II} m_1)} = \frac{m_2 S_b^2 + \left(1 - \frac{m_2}{M}\right) \bar{S}_w^2}{n_2^2 m_2 (c_2^I + c_1^{II} m_2)} \\
 &= \frac{\bar{S}_w^2}{m_1^2 n_1^2 c_1^{II}} = \frac{\bar{S}_w^2}{m_2^2 n_2^2 c_2^{II}}.
 \end{aligned}$$

Utilizing these conditions, we have

$$m^2 n^2 c^{II} \left[m S_b^2 + \left(1 - \frac{m}{M}\right) \bar{S}_w^2 \right] = n^2 m (c^I + c^{II} m) \bar{S}_w^2,$$

which can be written as

$$\begin{aligned}
 0 &= (c^I + c^{II} m) \bar{S}_w^2 - m c^{II} \left[m S_b^2 + \left(1 - \frac{m}{M}\right) \bar{S}_w^2 \right] \\
 &= (c^I + c^{II} m) M \bar{S}_w^2 - m c^{II} [M m S_b^2 + (M - m) \bar{S}_w^2] \\
 &= c^I M \bar{S}_w^2 + m c^{II} M \bar{S}_w^2 - m^2 c^{II} M S_b^2 - m c^{II} M \bar{S}_w^2 + m^2 c^{II} \bar{S}_w^2 \\
 &= c^I M \bar{S}_w^2 + m^2 c^{II} (\bar{S}_w^2 - M S_b^2).
 \end{aligned}$$

Hence,

$$m = \sqrt{\frac{c^I}{c^{II}} \frac{\bar{S}_w^2}{S_b^2 - \bar{S}_w^2 / M}}.$$

From the survey budget (3.1), the number of clusters is found to be

$$n = \frac{C_0}{2(c^I + m c^{II})} = \frac{C_0}{2\{c^I + [c^I c^{II} \bar{S}_w^2 / (S_b^2 - \bar{S}_w^2 / M)]^{1/2}\}}.$$

Plugging these expressions into (2.11) and using the equality relations (2.9), we obtain the variance of the estimator as

$$\begin{aligned}
 V_{e,t}[d] &= 2 \left(1 - \frac{n}{N}\right) \frac{S_b^2}{n} + 2 \left(1 - \frac{m}{M}\right) \frac{\bar{S}_w^2}{mn} \\
 &= 2 \left[1 - \frac{C_0}{2(c^I + m c^{II})N}\right] \frac{2(c^I + m c^{II}) S_b^2}{C_0} \\
 &\quad + 4 \left(1 - \frac{m}{M}\right) \frac{\bar{S}_w^2 (c^I + m c^{II})}{m C_0} \\
 &= 2 \left[\frac{2(c^I + m c^{II})}{C_0} - \frac{1}{N} \right] S_b^2 \\
 &\quad + 4 \left(\frac{1}{m} - \frac{1}{M} \right) \frac{\bar{S}_w^2 (c^I + m c^{II})}{C_0} \\
 &= \frac{4(c^I + m c^{II})}{C_0} \left[S_b^2 + \left(\frac{1}{m} - \frac{1}{M} \right) \bar{S}_w^2 \right] - \frac{2}{N} S_b^2 \\
 &= \frac{4 \left[c^I + \sqrt{c^I c^{II} \bar{S}_w^2 / (S_b^2 - \bar{S}_w^2 / M)} \right]}{C_0} \\
 &\quad \times \left[S_b^2 + \left(\sqrt{\frac{c^{II}}{c^I} \frac{S_b^2 - \bar{S}_w^2 / M}{\bar{S}_w^2}} - \frac{1}{M} \right) \bar{S}_w^2 \right] - \frac{2}{N} S_b^2.
 \end{aligned}$$

Proof of Proposition 5. The Lagrangian function of minimizing (2.13) subject to constraint (3.4) is

$$\begin{aligned}
 L(n, m_1, m_2, \lambda) &= 2 \left(1 - \frac{n}{N}\right) \frac{(1 - \rho^I) S_b^2}{n} \\
 &\quad + \left(1 - \frac{m_1}{M}\right) \frac{\bar{S}_{1w}^2}{n m_1} + \left(1 - \frac{m_2}{M}\right) \frac{\bar{S}_{2w}^2}{n m_2} \\
 &\quad - \lambda (c_{12}^I n + c_1^{II} n m_1 + c_2^{II} n m_2 - C_0).
 \end{aligned}$$

The first order conditions are:

$$\begin{aligned}
 \frac{\partial L}{\partial n} &= -2 \frac{(1 - \rho^I) S_b^2}{n^2} - \left(1 - \frac{m_1}{M}\right) \frac{\bar{S}_{1w}^2}{n^2 m_1} \\
 &\quad - \left(1 - \frac{m_2}{M}\right) \frac{\bar{S}_{2w}^2}{n^2 m_2} - \lambda (c_{12}^I + c_1^{II} m_1 + c_2^{II} m_2), \\
 \frac{\partial L}{\partial m_1} &= -\frac{\bar{S}_{1w}^2}{n m_1^2} - \lambda c_1^{II} n, \\
 \frac{\partial L}{\partial m_2} &= -\frac{\bar{S}_{2w}^2}{n m_2^2} - \lambda c_2^{II} n, \\
 \frac{\partial L}{\partial \lambda} &= c_{12}^I n + c_1^{II} n m_1 + c_2^{II} n m_2 - C_0 = 0.
 \end{aligned}$$

Expressing $-\lambda n$ from these conditions, one obtains:

$$\begin{aligned}
 -\lambda n &= \frac{\bar{S}_{1w}^2}{m_1^2 n c_1^{II}} = \frac{\bar{S}_{2w}^2}{m_2^2 n c_2^{II}} = 2(1 - \rho^I) \frac{S_b^2}{C_0} \\
 &\quad + \left(1 - \frac{m_1}{M}\right) \frac{\bar{S}_{1w}^2}{m_1 C_0} + \left(1 - \frac{m_2}{M}\right) \frac{\bar{S}_{2w}^2}{m_2 C_0}.
 \end{aligned}$$

Then

$$\begin{aligned}
 \frac{1}{m_2} &= \frac{1}{m_1} \sqrt{\frac{c_2^{II} \bar{S}_{1w}^2}{c_1^{II} \bar{S}_{2w}^2}} \equiv \frac{1}{\kappa m_1}, \\
 \frac{1}{m_1^2} \frac{(c_{12}^I + c_1^{II} m_1 + \kappa c_2^{II} m_1) \bar{S}_{1w}^2}{c_1^{II}} &= 2(1 - \rho^I) S_b^2 \\
 &\quad + \left(\frac{1}{m_1} - \frac{1}{M} \right) \bar{S}_{1w}^2 + \left(\frac{1}{\kappa m_1} - \frac{1}{\kappa M} \right) \bar{S}_{2w}^2, \\
 0 &= [2(1 - \rho^I) S_b^2 \kappa c_1^{II} - \bar{S}_{1w}^2 \kappa c_1^{II} / M - \bar{S}_{2w}^2 c_1^{II} / M] m_1^2 \\
 &\quad + [\bar{S}_{1w}^2 \kappa c_1^{II} + \bar{S}_{2w}^2 c_1^{II} - c_1^{II} \bar{S}_{1w}^2 \kappa - \kappa^2 c_2^{II} \bar{S}_{1w}^2] m_1 - c_{12}^I \bar{S}_{1w}^2 \kappa, \\
 D &= [\bar{S}_{1w}^2 \kappa c_1^{II} + \bar{S}_{2w}^2 c_1^{II} - c_1^{II} \bar{S}_{1w}^2 \kappa - \kappa^2 c_2^{II} \bar{S}_{1w}^2]^2 \\
 &\quad + 4[2(1 - \rho^I) S_b^2 \kappa - \bar{S}_{1w}^2 \kappa / M - \bar{S}_{2w}^2 / M] c_1^{II} c_{12}^I \bar{S}_{1w}^2 \geq 0, \\
 m_1 &= \frac{c_1^{II} \bar{S}_{1w}^2 \kappa + \kappa^2 c_2^{II} \bar{S}_{1w}^2 - \bar{S}_{1w}^2 \kappa c_1^{II} - \bar{S}_{2w}^2 c_1^{II} \pm \sqrt{D}}{4(1 - \rho^I) S_b^2 \kappa c_1^{II} - 2 \bar{S}_{1w}^2 \kappa c_1^{II} / M - 2 \bar{S}_{2w}^2 c_1^{II} / M}.
 \end{aligned}$$

The solution with $-\sqrt{D}$ leads to a negative value of m_1 , and must be discarded.

The remaining design characteristics are

$$m_2 = \kappa m_1, \quad n = \frac{C_0}{c_{12}^I + m_1 c_1^{II} + m_2 c_2^{II}}, \quad \kappa = \sqrt{\frac{c_1^{II} \bar{S}_{2w}^2}{c_2^{II} \bar{S}_{1w}^2}}.$$

The variance of the difference estimator can be found using (2.15).

Under symmetric conditions, $\kappa = 1$, and

$$D = 4[2(1 - \rho^I)S_b^2 - 2\bar{S}_w^2/M]c_{12}^I\bar{S}_w^2$$

is non-negative unless the expression in the square brackets is negative (which can only happen when ρ^I is large and M is small. In that case, a corner solution $m = M$ is realized). Furthermore,

$$m = m_1 = m_2 = \sqrt{\frac{\bar{S}_w^2 c_{12}^I}{2[(1 - \rho^I)S_b^2 - \bar{S}_w^2/M]c_{12}^I}},$$

$$n = \frac{C_0}{c_{12}^I + 2mc^{\Pi}} = \frac{C_0}{c_{12}^I + \sqrt{\frac{2\bar{S}_w^2 c_{12}^I c^{\Pi}}{(1 - \rho^I)S_b^2 - \bar{S}_w^2/M}}},$$

$V_{e.o}[d]$

$$\begin{aligned} &= 2\left(1 - \frac{n}{N}\right)\frac{(1 - \rho^I)S_b^2}{n} + 2\left(1 - \frac{m}{M}\right)\frac{(1 - \rho^{\Pi})\bar{S}_w^2}{nm} \\ &= \frac{2}{n}\left[(1 - \rho^I)S_b^2 + 2\left(1 - \frac{m}{M}\right)\frac{(1 - \rho^{\Pi})\bar{S}_w^2}{m}\right] - \frac{2(1 - \rho^I)S_b^2}{N} \\ &= \frac{2}{C_0}(c_{12}^I + 2mc^{\Pi}) \\ &\quad \times \left[(1 - \rho^I)S_b^2 + 2\left(\frac{1}{m} - \frac{1}{M}\right)(1 - \rho^{\Pi})\bar{S}_w^2\right] - \frac{2(1 - \rho^I)S_b^2}{N} \\ &\quad - \frac{2}{C_0}(c_{12}^I + 2mc^{\Pi}) \\ &\quad \times \left[(1 - \rho^I)S_b^2 + \frac{2}{m}(1 - \rho^{\Pi})\bar{S}_w^2 - \frac{2}{M}(1 - \rho^{\Pi})\bar{S}_w^2\right] - \frac{2(1 - \rho^I)S_b^2}{N} \\ &= \frac{2}{C_0}\left\{c_{12}^I(1 - \rho^I)S_b^2 + 2(1 - \rho^{\Pi})\bar{S}_w^2\left[2c^{\Pi} - \frac{c_{12}^I}{M}\right]\right. \\ &\quad \left.+ \frac{2}{m}c_{12}^I(1 - \rho^{\Pi})\bar{S}_w^2 + 2mc^{\Pi}\left[(1 - \rho^I)S_b^2 - \frac{2}{M}(1 - \rho^{\Pi})\bar{S}_w^2\right]\right\} \\ &\quad - \frac{2(1 - \rho^I)S_b^2}{N} \\ &= \frac{2}{C_0}\left\{c_{12}^I(1 - \rho^I)S_b^2 + 2(1 - \rho^{\Pi})\bar{S}_w^2\left[2c^{\Pi} - \frac{c_{12}^I}{M}\right]\right. \\ &\quad \left.+ 2(1 - \rho^{\Pi})\sqrt{2[(1 - \rho^I)S_b^2 - \bar{S}_w^2/M]}\bar{S}_w^2 c_{12}^I\right. \\ &\quad \left.+ \sqrt{\frac{2\bar{S}_w^2 c_{12}^I c^{\Pi}}{(1 - \rho^I)S_b^2 - \bar{S}_w^2/M}}\left[(1 - \rho^I)S_b^2 - \frac{2}{M}(1 - \rho^{\Pi})\bar{S}_w^2\right]\right\} \\ &\quad - \frac{2(1 - \rho^I)S_b^2}{N}. \end{aligned}$$

Proof of Proposition 6. The Lagrangian function of minimizing (2.15) subject to constraint (3.6) is

$$L(n, m, \lambda) = 2\left(1 - \frac{n}{N}\right)\frac{(1 - \rho^I)S_b^2}{n} + 2\left(1 - \frac{m}{M}\right)\frac{(1 - \rho^{\Pi})\bar{S}_w^2}{nm} - \lambda(c_{12}^I n + c_{12}^{\Pi} nm - C_0).$$

The first order conditions are:

$$\begin{aligned} 0 &= \frac{\partial L}{\partial n} = -2\frac{(1 - \rho^I)S_b^2}{n^2} - 2\left(1 - \frac{m}{M}\right)\frac{(1 - \rho^{\Pi})\bar{S}_w^2}{n^2 m} - \lambda(c_{12}^I + c_{12}^{\Pi} m), \\ 0 &= \frac{\partial L}{\partial m} = -2\frac{(1 - \rho^{\Pi})\bar{S}_w^2}{nm^2} - \lambda c_{12}^{\Pi} n, \\ 0 &= \frac{\partial L}{\partial \lambda} = c_{12}^I n + c_{12}^{\Pi} nm - C_0. \end{aligned}$$

Expressing $-\lambda n^2$ from these conditions, one obtains:

$$\begin{aligned} -\lambda n^2/2 &= \frac{(1 - \rho^I)S_b^2}{c_{12}^I + c_{12}^{\Pi} m} + \left(1 - \frac{m}{M}\right)\frac{(1 - \rho^{\Pi})\bar{S}_w^2}{m(c_{12}^I + c_{12}^{\Pi} m)} \\ &= \frac{(1 - \rho^{\Pi})\bar{S}_w^2}{m^2 c_{12}^{\Pi}}. \end{aligned}$$

Hence,

$$\begin{aligned} (1 - \rho^I)S_b^2 M m^2 c_{12}^{\Pi} + (M - m)(1 - \rho^{\Pi})\bar{S}_w^2 m c_{12}^{\Pi} \\ - (1 - \rho^{\Pi})M\bar{S}_w^2(c_{12}^I + c_{12}^{\Pi} m) &= 0, \\ [(1 - \rho^I)S_b^2 M c_{12}^{\Pi} - (1 - \rho^{\Pi})\bar{S}_w^2 c_{12}^{\Pi}]m^2 \\ - [(1 - \rho^{\Pi})M\bar{S}_w^2 c_{12}^I] &= 0, \\ m &= \sqrt{\frac{(1 - \rho^{\Pi})M\bar{S}_w^2 c_{12}^I}{[(1 - \rho^I)S_b^2 M - (1 - \rho^{\Pi})\bar{S}_w^2]c_{12}^{\Pi}}} \\ &= \sqrt{\frac{c_{12}^I}{c_{12}^{\Pi}}\frac{(1 - \rho^{\Pi})\bar{S}_w^2}{(1 - \rho^I)S_b^2 - (1 - \rho^{\Pi})\bar{S}_w^2/M}}. \end{aligned}$$

From the survey budget (3.6),

$$n = \frac{C_0}{c_{12}^I + c_{12}^{\Pi} m} = \frac{C_0}{c_{12}^I + \sqrt{\frac{(1 - \rho^{\Pi})\bar{S}_w^2 c_{12}^I c_{12}^{\Pi}}{(1 - \rho^I)S_b^2 - (1 - \rho^{\Pi})\bar{S}_w^2/M}}}.$$

Finally, the variance of the difference estimator is

$$\begin{aligned}
V_{e,o}[d] &= \frac{2}{C_0} \left(c_{12}^I + \sqrt{\frac{(1-\rho^{II})\bar{S}_w^2 c_{12}^I c_{12}^{II}}{(1-\rho^I)S_b^2 - (1-\rho^{II})\bar{S}_w^2 / M}} \right) \\
&\times \left[(1-\rho^I)S_b^2 \right. \\
&\quad \left. + \left(\sqrt{\frac{c_{12}^{II}(1-\rho^I)S_b^2 - (1-\rho^{II})\bar{S}_w^2 / M}{(1-\rho^{II})\bar{S}_w^2}} - \frac{1}{M} \right) (1-\rho^{II})\bar{S}_w^2 \right] \\
&\quad - \frac{2(1-\rho^I)S_b^2}{N} \\
&= \frac{2}{C_0} \left\{ (1-\rho^I)S_b^2 c_{12}^I \right. \\
&\quad + (1-\rho^{II})\bar{S}_w^2 \sqrt{c_{12}^I c_{12}^{II} \frac{(1-\rho^I)S_b^2 - (1-\rho^{II})\bar{S}_w^2 / M}{(1-\rho^{II})\bar{S}_w^2}} \\
&\quad + \left[(1-\rho^I)S_b^2 - \frac{1}{M}(1-\rho^{II})\bar{S}_w^2 \right] \\
&\quad \times \sqrt{\frac{(1-\rho^{II})\bar{S}_w^2 c_{12}^I c_{12}^{II}}{(1-\rho^I)S_b^2 - (1-\rho^{II})\bar{S}_w^2 / M}} \\
&\quad \left. + (1-\rho^{II})\bar{S}_w^2 \left(c_{12}^{II} - \frac{c_{12}^I}{M} \right) \right\} \\
&\quad - \frac{2(1-\rho^I)S_b^2}{N}.
\end{aligned}$$

Proof of Proposition 7. Ignoring the finite population correcting terms of the order $O(N^{-1})$ and $O(M^{-1})$, equation (3.3) can be written as:

$$\begin{aligned}
V_{e,i}[d] &\approx \frac{4(c^I + \sqrt{c^I c^{II} \bar{S}_w^2 / S_b^2})}{C_0} \left[S_b^2 + \left(\sqrt{\frac{c^{II} \bar{S}_w^2}{c^I}} S_b^2 \right) \right] \\
&= \frac{4}{C_0} (c^I S_b^2 + c^{II} \bar{S}_w^2 + 2\sqrt{c^I c^{II} S_b^2 \bar{S}_w^2}) \\
&= \frac{4}{C_0} (\sqrt{c^I S_b^2} + \sqrt{c^{II} \bar{S}_w^2})^2.
\end{aligned}$$

Likewise, equation (3.8) can be written as

$$\begin{aligned}
V_{e,o}[d] &\approx \frac{2}{C_0} \left[(1-\rho^I)S_b^2 c_{12}^I \right. \\
&\quad \left. + 2\sqrt{c_{12}^I c_{12}^{II} (1-\rho^I)S_b^2 (1-\rho^{II})\bar{S}_w^2} + (1-\rho^{II})\bar{S}_w^2 c_{12}^{II} \right] \\
&= \frac{2}{C_0} \left[\sqrt{(1-\rho^I)S_b^2 c_{12}^I} + \sqrt{(1-\rho^{II})\bar{S}_w^2 c_{12}^{II}} \right]^2.
\end{aligned}$$

The statement of Proposition 7 follows immediately from these two expressions.

References

- Binder, D.A., and Hidirolou, M.A. (1988). Sampling in time. In *Handbook of Statistics*, (Eds., P.R. Krishnaiah and C.R. Rao), North Holland, Amsterdam, 6, 187-211.
- Cochran, W.G. (1977). *Sampling Techniques*, 3rd Ed., New York: John Wiley & Sons, Inc.
- Eckler, A.R. (1955). Rotation sampling. *Annals of Mathematical Statistics*, 26(4), 664-685.
- Ernst, L.R. (1999). The maximization and minimization of sample overlap problems: A half century of results. Technical report, U.S. Bureau of Labor Statistics.
- Fuller, W.A. (1999). Environmental surveys over time. *Journal of Agricultural, Biological and Environmental Statistics*, 4(4), 331-345.
- Groves, R.M. (1989). *Survey Errors and Survey Costs*. New York: John Wiley & Sons, Inc.
- Hansen, M., Hurwitz, W.N. and Madow, W.G. (1953). *Sample Survey Methods and Theory*. New York: John Wiley & Sons, Inc.
- Kish, L. (1995). *Survey Sampling*, 3rd Ed., New York: John Wiley & Sons, Inc.
- Lehtonen, R., and Pahkinen, E. (2004). *Practical Methods for Design and Analysis of Complex Surveys, Statistics in Practice*, 2nd Ed., New York: John Wiley & Sons, Inc.
- Mas-Colell, A., Whinston, M.D. and Green, J.R. (1995). *Microeconomic Theory*, Oxford University Press, Oxford, UK.
- McDonald, T.L. (2003). Review of environmental monitoring methods: Survey designs. *Environmental Monitoring and Assessment*, 85, 277-292.
- Neyman, J. (1938). Contribution to the theory of sampling human populations. *The Journal of the American Statistical Association*, 33, 101-116.
- Patterson, H.D. (1950). Sampling on successive occasions with partial replacement of units. *Journal of the Royal Statistical Society, Series B*, 12(2), 241-255.
- Rao, J.N.K., and Graham, J.E. (1964). Rotation designs for sampling on repeated occasions. *Journal of the American Statistical Association*, 59(306), 492-509.
- Särndal, C.-E., Swensson, B. and Wretman, J. (1992). *Model Assisted Survey Sampling*, New York: Springer.

- Scott, C.T. (1998). Sampling methods for estimating change in forest resources. *Ecological Applications*, 8(2), 228-233.
- Thompson, M.E. (1997). Theory of Sample Surveys. *Monographs on Statistics and Applied Probability*, New York: Chapman & Hall/CRC, 74.
- Thompson, S.K. (1992). *Sampling*, New York: John Wiley & Sons, Inc.
- Wolter, K.M. (2007). *Introduction to Variance Estimation*, 2nd Ed., New York: Springer.

On the efficiency of randomized probability proportional to size sampling

Paul Knottnerus¹

Abstract

This paper examines the efficiency of the Horvitz-Thompson estimator from a systematic probability proportional to size (PPS) sample drawn from a randomly ordered list. In particular, the efficiency is compared with that of an ordinary ratio estimator. The theoretical results are confirmed empirically with a simulation study using Dutch data from the Producer Price Index.

Key Words: Horvitz-Thompson estimator; Producer Price Index; Ratio estimator; Sampling autocorrelation coefficient.

1. Introduction

When the study variable y in a population of N units is more or less proportional to a size variable x , one may use the ratio estimator from a simple random sample of size n without replacement (SRS). An alternative estimator in such a situation is the Horvitz-Thompson (HT) estimator in combination with a systematic probability proportional to size sample from a randomly ordered list, henceforth called a randomized PPS sample.

In recent years several authors investigated variance estimation procedures for the HT estimator from a randomized PPS sample. See, among others, Brewer and Donadio (2003), Cumberland and Royall (1981), Deville (1999), Knottnerus (2003), Kott (1988 and 2005), Rosén (1997) and Stehman and Overton (1994). For a comparison between the efficiencies of the ratio estimator and the randomized PPS estimator, the reader is referred to Foreman and Brewer (1971), Cochran (1977) and the references given therein. A drawback of these comparisons is that finite populations corrections are ignored. Hartley and Rao (1962) take the finite population correction into account but without an explicit formula for the efficiency. Elaborating on the results of Gabler (1984), Qualité (2008) shows that the related HT estimator from a rejective Poisson sample of size n is more efficient than the Hansen-Hurwitz estimator for a sampling scheme with replacement. No formula for the increased efficiency is given, however.

The main aim of this paper is to derive formulas for the efficiency of the randomized PPS estimator relative to the ratio estimator. To this end, we present a simple formula for the change in the sample size required to maintain the same variance when a randomized PPS estimator is replaced by a ratio estimator. From the design based point of view these formulas are valid when $n = o(N)$ as $N \rightarrow \infty$. This condition suggests that the finite population correction can be neglected for this kind of sampling design. Surprisingly, as we will see in an example in section 4, the randomized PPS sampling can reduce variance by more than 30% compared

to PPS sampling *with* replacement even when the sampling fraction n/N is much smaller than 30%; see also Kott (2005, page 436). Furthermore, the formulas remain appropriate from a model assisted point of view when n and N are of the same order, provided that N is large and that the hypothetical model for the observations Y_i ($i = 1, \dots, N$) satisfies mild conditions.

The outline of the paper is as follows. Section 2 describes an alternative expression for the variance of the HT estimator based on the sampling autocorrelation coefficient. The corresponding variance estimator for randomized PPS sampling is shown to be nonnegative with probability 1. Section 3 presents the formulas for the efficiency of the randomized PPS estimator relative to the ratio estimator for various data patterns often met in practice. Section 4 features an example with data on the Producer Price Index in The Netherlands illustrating the substantial efficiency gains obtainable in practice. A counterexample shows that randomized PPS sampling is not *always* advantageous. The paper concludes with a summary.

2. An alternative variance expression for randomized PPS sampling

Consider a population $U = \{1, \dots, N\}$, and let s be a sample of fixed size n drawn from U without replacement according to a given sampling design with first order inclusion probabilities π_i and second order inclusion probabilities π_{ij} ($i, j = 1, \dots, N$). The HT estimator of the population total, $Y = \sum_{i \in U} Y_i$, is defined by $\hat{Y}_{HT} = \sum_{i \in s} Y_i / \pi_i$. Suppose there is a measure of relative size X_i (i.e., $X = \sum_{i \in U} X_i = 1$) such that all $X_i \leq 1/n$. In fact, it is assumed here that units with $X_i > 1/n$ are put together in a separate certainty-stratum. When the π_i are proportional to these size measures, $\pi_i = nX_i$. Defining $Z_i = Y_i/X_i$, we can write Y as a weighted mean of the Z_i , that is, $Y = \mu_z = \sum_{i \in U} X_i Z_i$. Likewise, we can write the HT estimator of Y in randomized PPS sampling as $\hat{Y}_{HT} = \hat{Y}_{PPS} = \bar{z}_s$, where \bar{z}_s is sample mean of the Z_i .

1. Paul Knottnerus, Statistics Netherlands, PO Box 24500, 2490 HA The Hague, The Netherlands. E-mail: pkts@cbs.nl.

The variance of the randomized PPS estimator \hat{Y}_{PPS} is

$$\text{var}(\hat{Y}_{PPS}) = \frac{1}{n^2} \sum_{i \in U} \sum_{j \in U} (\pi_{ij} - \pi_i \pi_j) Z_i Z_j \quad (1)$$

$$= -\frac{1}{2n^2} \sum_{i \in U'} \sum_{j \in U'} (\pi_{ij} - \pi_i \pi_j) (Z_i - Z_j)^2 \quad (2)$$

with $\pi_{ii} = \pi_i$. The former is attributed to Horvitz and Thompson (1952) and the latter is due to Sen (1953) and Yates and Grundy (1953). The following alternative expression for the variance is more convenient for our purposes:

$$\text{var}(\hat{Y}_{PPS}) = \text{var}(\bar{z}_s) = \{1 + (n-1)\rho_z\} \frac{\sigma_z^2}{n}, \quad (3)$$

where $\sigma_z^2 = \sum_{i \in U} X_i (Z_i - \mu_z)^2$, and

$$\rho_z = \sum_{i \in U} \sum_{\substack{j \in U \\ j \neq i}} \frac{\pi_{ij}}{n(n-1)} \left(\frac{Z_i - \mu_z}{\sigma_z} \right) \left(\frac{Z_j - \mu_z}{\sigma_z} \right). \quad (4)$$

For a proof of (3), see Knottnerus (2003, page 103). Note that σ_z^2/n would have been the variance if the sample had been drawn with replacement with drawing probabilities X_i .

The sampling autocorrelation coefficient ρ_z in (4) is a generalization of the more familiar intraclass correlation coefficient ρ in systematic sampling with equal probabilities; see, for instance, Cochran (1977, pages 209 and 240) and Särndal, Swensson and Wretman (1992, page 79). Note that ρ_z is a fixed population parameter. The phrase *sampling autocorrelation* is used because ρ_z refers to the autocorrelation between two randomly chosen observations, say z_{s1} and z_{s2} , from s . Consequently, the value of ρ_z depends on the sampling design. In particular, when sampling with replacement, $\rho_z = 0$, while under SRS sampling, $\rho_z = -1/(N-1)$.

Although exact expressions for the π_{ij} under randomized PPS sampling are available, they can be cumbersome when N is large. For an exact expression, see Connor (1966) and for a modification Hidiroglou and Gray (1980). Here we use an approximation proposed by Knottnerus (2003, page 197):

$$\pi_{ijK} = n(n-1) \frac{X_i X_j (1 - X_i - X_j)}{\gamma (1 - 2X_i)(1 - 2X_j)} \quad (5)$$

$$\gamma = \frac{1}{2} + \frac{1}{2} \sum_{i \in U} \frac{X_i}{1 - 2X_i}.$$

These π_{ijK} have been shown to satisfy the second-order restrictions for the π_{ij} :

$$\sum_{i, j \in U' (j \neq i)} \pi_{ij} = n(n-1),$$

and

$$\sum_{j \in U' (j \neq i)} \pi_{ij} = (n-1)\pi_i.$$

Furthermore, (5) is correct for SRS sampling for any $n \leq N$, while π_{ijK} coincide with the π_{ijBD} from the special designs proposed by Brewer (1963a) and Durbin (1967) for PPS samples with $n = 2$. Moreover, the π_{ijK} in (5) can be written in factorized form as proposed by Brewer and Donadio (2003). That is,

$$\pi_{ijK} = \pi_i \pi_j (c_i + c_j) / 2, \quad (6)$$

and

$$c_i = (n-1)/n\gamma(1-2X_i).$$

An implication of approximation (5) is that $\pi_{ijK}/n(n-1)$ does not depend on n . Hence, the corresponding approximation of ρ_z does not depend on n (recall we have assumed that every $X_i < 1/n$).

This nondependence on n would also result had we used the approximation proposed by Hartley and Rao (1962) for randomized PPS sampling:

$$\begin{aligned} \pi_{ijHR} &= n(n-1) X_i X_j \\ &\quad \{1 + X_i + X_j - \mu_x + 2(X_i^2 + X_j^2 + X_i X_j) \\ &\quad - 3\mu_x (X_i + X_j - \mu_x - 2\sum_{i \in U} X_i^3)\}, \end{aligned} \quad (7)$$

where $\mu_x = \sum_{i \in U} X_i^2$ (recall $\mu_z = \sum_{i \in U} X_i Z_i$). Obviously, $\pi_{ijHR}/n(n-1)$ does not depend on n . At the time Hartley and Rao assumed that $n = O(1)$ as $N \rightarrow \infty$. In addition, referring to a private conversation with J.N.K. Rao, Thompson and Wu (2008) state that approximation (7) is valid when $n = o(N)$ as $N \rightarrow \infty$. For an example that (5) and (7) can not be used for any n and N , see Appendix A.

Since both (5) and (7) lead to approximations for ρ_z in randomized PPS sampling that are $\rho_z \{1 + o(1)\}$ as $N \rightarrow \infty$ with $n = o(N)$, (5) can be used for calculating ρ_z in practice when $n \ll N$ and N is large. For ease of the exposition, it is assumed here that there is a positive constant c such that $\rho_z < -c/N$. See also Kott (2005, page 436) who discusses estimating the variance under PPS sampling when $n = O(N^{2/3})$.

Suppose $\gamma = 1 + \mu_x + O(1/N^2)$ and $\mu_x = O(1/N)$ (which follow from the conditions of Theorem 1 below). It is not hard to see that, after dropping $O(1/nN)$ terms, c_i in (6) is identical with $c_{iHR} = (n-1)/\{n(1 + \mu_x - 2X_i)\}$. The latter expression is equation (11) of Brewer and Donadio, which is based on π_{ijHR} in (7).

The approach proposed here is somewhat different from Knottnerus (2003). First, rewrite (5) as

$$\pi_{ijK} = n(n-1) \frac{X_i X_j}{\gamma} \left(\frac{1/2}{1 - 2X_i} + \frac{1/2}{1 - 2X_j} \right). \quad (8)$$

Substituting (8) into (4), we obtain a new, simple approximation for ρ_z :

$$\begin{aligned}\rho_z &= \sum_{i \in U} \sum_{\substack{j \in U \\ j \neq i}} \frac{X_i X_j}{\gamma} \left(\frac{1/2}{1-2X_i} + \frac{1/2}{1-2X_j} \right) \left(\frac{Z_i - Y}{\sigma_z} \right) \left(\frac{Z_j - Y}{\sigma_z} \right) \\ &= \sum_{i \in U} \sum_{\substack{j \in U \\ j \neq i}} \frac{X_i X_j}{\gamma} \left(\frac{1}{1-2X_i} \right) \left(\frac{Z_i - Y}{\sigma_z} \right) \left(\frac{Z_j - Y}{\sigma_z} \right) \\ &= 0 - \sum_{i \in U} \frac{X_i^2}{\gamma(1-2X_i)} \left(\frac{Z_i - Y}{\sigma_z} \right)^2.\end{aligned}\quad (9)$$

In the second line, we used the equality $\sum_{i,j} m_{ij} v_i = \sum_{i,j} m_{ij} v_j$ when $m_{ij} = m_{ji}$. In the last line, we used $\sum_{j \in U} X_j (Z_j - Y) = 0$.

Next, let \bar{X} denote the population mean of X_1, \dots, X_N and define σ_x^2 and V_x^2 by

$$\sigma_x^2 = \sum_{i \in U} X_i (X_i - \mu_x)^2,$$

and

$$V_x^2 = \sum_{i \in U} (X_i - \bar{X})^2 / N,$$

respectively. In the following theorem (9) is further simplified.

Theorem 1. Suppose that $(Z_i - Y)/\sigma_z = O(1)$ as $N \rightarrow \infty$ and that there are positive constants c and C such that $V_x/\bar{X} < c$, $\sigma_x/\mu_x < c$ and $0 < X_i < C < 1/2$. Then, for large N and $n \ll N$,

$$\rho_z = - \frac{\sum_{i \in U} X_i^2 (Z_i - Y)^2}{\sum_{i \in U} X_i (Z_i - Y)^2} \left\{ 1 + O\left(\frac{1}{N}\right) \right\} + O\left(\frac{1}{N^2}\right). \quad (10)$$

Proof. Because $\bar{X} = 1/N$, it follows from the above assumptions that the weighted mean $\mu_x [= \sum X_i^2 = N(V_x^2 + \bar{X}^2)]$ is of order $1/N$ and hence, $\sigma_x = O(1/N)$. Because $(1-2X_i)^{-1} = 1 + 2X_i + O(X_i^2)$ for $0 < X_i < C < 1/2$, ρ_z from (9) can be written for $N \rightarrow \infty$ as

$$\rho_z = - \sum_{i \in U} \frac{X_i^2}{\gamma} \left(\frac{Z_i - Y}{\sigma_z} \right)^2 + \frac{1}{\gamma} O\left(\sum_{i \in U} X_i^3 \right),$$

where $\sum_{i \in U} X_i^3 = \sigma_x^2 + \mu_x^2 = O(N^{-2})$, and

$$\begin{aligned}\gamma &= \frac{1}{2} + \frac{1}{2} \sum_{i \in U} X_i \{1 + 2X_i + O(X_i^2)\} \\ &= 1 + \mu_x + O\left(\frac{1}{N^2}\right) = 1 + O\left(\frac{1}{N}\right),\end{aligned}$$

from which (10) follows. This concludes the proof.

Substituting (10) into (3), we get

$$\begin{aligned}\text{var}(\hat{Y}_{\text{PPS}}) &= \frac{\sigma_z^2}{n} - \frac{n-1}{n} \sum_{i \in U} X_i^2 (Z_i - Y)^2 \\ &= \frac{1}{n} \sum_{i \in U} X_i \{1 - (n-1)X_i\} (Z_i - Y)^2,\end{aligned}\quad (11)$$

which is also given by Hartley and Rao (1962). It is noteworthy that approximation (10) also follows directly from substituting the simple approximation $\pi_{ijAP} = n(n-1)X_i X_j$ into (4). Likewise, use of π_{ijHR} leads to an expression almost similar to (9) and hence to (10). In addition, direct use of π_{ijAP} in (1) or (2) for the SRS case with $X_i = X_j = 1/N$ may lead to errors of more than 100% for populations with $\bar{Y} = V_y^2$; see Knottnerus (2003, pages 274-6). Hence, (1) and (2) are more sensitive to small errors in the π_{ij} than (3) and (4). Furthermore, note that when n is so small that $|n\rho_z| \ll 1$, we may set $\rho_z = 0$ yielding the with-replacement variance formula of Hansen and Hurwitz (1943).

In order to estimate (3) using ρ_z , denote, as before, a randomly chosen observation from s by z_{s1} . Then we have

$$\begin{aligned}\sigma_z^2 &= \text{var}(z_{s1}) = \text{var}\{E(z_{s1}|s)\} + E\{\text{var}(z_{s1}|s)\} \\ &= \text{var}(\bar{z}_s) + E\left(\frac{n-1}{n} s^2\right),\end{aligned}$$

where

$$s^2 = \frac{1}{n-1} \sum_{i \in s} (Z_i - \bar{z}_s)^2.$$

Now from (3), it is seen that $s_z^2/(1-\rho_z)$ is an unbiased estimator for σ_z^2 . When ρ_z is very small, the term $(1-\rho_z)$ can be neglected. When n is sufficiently large, the ratio ρ_z from (9) can be estimated by

$$\hat{\rho}_{z9} = - \frac{\sum_{i \in s} X_i (Z_i - \bar{z}_s)^2 / \hat{\gamma} (1 - 2X_i)}{\sum_{i \in s} (Z_i - \bar{z}_s)^2},$$

where

$$\hat{\gamma} = \frac{1}{2} + \frac{1}{2n} \sum_{i \in s} \frac{1}{1 - 2X_i}.$$

Because $\hat{\gamma} \geq 1$ and $X_i \leq 1/n$, we have $\hat{\rho}_{z9} \geq -1/(n-2)$. For the bias of an estimated ratio when n is small, see Cochran (1977, page 160).

In a similar manner ρ_z from (10) can be estimated by

$$\hat{\rho}_{z10} = - \frac{\sum_{i \in s} X_i (Z_i - \bar{z}_s)^2}{\sum_{i \in s} (Z_i - \bar{z}_s)^2} \geq \frac{-1}{n} > \frac{-1}{n-1}.$$

Hence, replacing σ_z^2 and ρ_z in (3) by $s_z^2/(1-\hat{\rho}_{z10})$ and $\hat{\rho}_{z10}$, respectively, leads to a nonnegative variance estimator

with probability 1. This also holds for $\hat{\rho}_{z9}$ when all $X_i \leq 1/(n+1)$. The estimator for $\text{var}(\hat{Y}_{\text{PPS}})$ thus obtained becomes

$$\hat{\text{var}}_{\rho}(\hat{Y}_{\text{PPS}}) = \frac{\{1 + (n-1)\hat{\rho}_{z9}\}s_z^2}{n(1 - \hat{\rho}_{z9})}.$$

Moreover, for moderate values of N , estimator $\hat{\rho}_{z9}$ has probably better properties than $\hat{\rho}_{z10}$ because the π_{ijk} underlying (9) satisfy exactly the second-order restrictions irrespective of the values of n and N .

3. Efficiency of \hat{Y}_{PPS} for large n and N

3.1 Efficiency formulas

Because $X = 1$, the ratio estimator for Y becomes

$$\hat{Y}_R = \frac{\bar{y}_s}{\bar{x}_s} = \frac{\sum_{i \in s} X_i Z_i}{\sum_{i \in s} X_i}.$$

For sufficiently large n the commonly used approximation for its variance is

$$\text{var}(\hat{Y}_R) = \frac{N(N-n)}{n(N-1)} \sum_{i \in U} X_i^2 (Z_i - Y)^2. \quad (12)$$

From (3) and (12) it can be seen that the efficiency of \hat{Y}_{PPS} relative to \hat{Y}_R can be written as

$$\text{Eff}_{P/R} = \frac{\text{var}(\hat{Y}_R)}{\text{var}(\hat{Y}_{\text{PPS}})} = \frac{(N-n) \sum_{i \in U} X_i^2 (Z_i - Y)^2}{\{1 + (n-1)\rho_z\}\sigma_z^2}, \quad (13)$$

assuming $N/(N-1) \approx 1$. Combining (10) and (13) gives

$$\text{Eff}_{P/R} = \frac{-(N-n)\rho_z}{1 + (n-1)\rho_z}. \quad (14)$$

Now suppose that the observations Y_i satisfy the model:

$$Y_i = \mu X_i + \varepsilon_i, \quad (15)$$

with $E(\varepsilon_i) = 0$, $E(\varepsilon_i^2) = \sigma^2 X_i^\delta$, and $E(\varepsilon_i \varepsilon_j) = 0$ ($i \neq j$). Consequently, for the Z_i we have $Z_i = \mu + u_i$ with $E(u_i) = 0$, $E(u_i^2) = \sigma^2 X_i^{\delta-2}$, and $E(u_i u_j) = 0$ ($i \neq j$). According to Kott (1988), δ often lies between 1 and 2. See also Brewer (1963b). Brewer and Donadio (2003) showed that by assuming a model like (15), (7) and hence (10) and (14) hold when n and N are of the same order as $N \rightarrow \infty$. Furthermore, for sufficiently large N we can replace Y as well as the numerator and denominator in (10) by their model expectations. This yields

$$\rho_z = -\frac{\sum_{i \in U} X_i^\delta}{\sum_{i \in U} X_i^{\delta-1}}. \quad (16)$$

In the next subsections we look more closely at the relationship between δ and the efficiency of \hat{Y}_{PPS} .

3.2 Efficiency of \hat{Y}_{PPS} when $\delta = 2$

For $\delta = 2$, (16) gives $\rho_z = -\sum_{i \in U} X_i^2 = -\mu_x$, which can also be written as

$$\rho_z = -\frac{1}{N}(1 + CV_x^2), \quad (17)$$

because

$$\frac{1}{N} \sum_{i \in U} X_i^2 = V_x^2 + \bar{X}^2 = \bar{X}^2(1 + CV_x^2),$$

where $\bar{X} = 1/N$ and $CV_x = V_x/\bar{X}$ is the coefficient of variation of the X_i . Substituting (17) into (14) gives

$$\text{Eff}_{P/R} = \frac{(N-n)(1 + CV_x^2)}{N - (n-1)(1 + CV_x^2)}.$$

Hence, for $\delta = 2$, the efficiency of the randomized PPS sample is high when the variability among the X_i is high. When $CV_x = 0$, randomized PPS sampling amounts to SRS sampling and obviously, $\text{Eff}_{P/R} = 1$ assuming $(N-n+1) \approx (N-n)$; note that this assumption holds when N is sufficiently large and $n/N < f_0 < 1$.

Observe that substituting $n = n_{\text{PPS}}(1 + CV_x^2)$ into (12) leads to about the same outcome as (3) and (10) with n_{PPS} instead of n . Hence, when $CV_x = 1.5$, randomized PPS sampling with sample size $n_{\text{PPS}} = 100$ is as efficient as the ratio estimator from an SRS sample of size $n_{\text{SRS}} = 325$. More generally, assuming that $(n-1)/n \approx 1$, it is seen from (3), (10), and (12) that a ratio estimator from an SRS sample of size n_{SRS} is as efficient as a PPS sample of size n_{PPS} when

$$n_{\text{SRS}} = -n_{\text{PPS}} \rho_z N. \quad (18)$$

3.3 Efficiency of \hat{Y}_{PPS} for $\delta < 1$ vs $\delta \geq 1$

Another special case is $\delta = 1$. From (16), $\rho_z = -1/N$ when $\delta = 1$. Subsequently, it follows from (14) that under model (15) $\text{Eff}_{P/R} = 1 + O(N^{-1})$, provided that $n/N < f_0 < 1$ as $N \rightarrow \infty$ irrespective of the value of CV_x . Furthermore, it can be shown that $\text{Eff}_{P/R}$ is an increasing function of δ . This is proven below in Lemma 1. Hence, for $\delta < 1$ the randomized PPS estimator is less efficient than the ratio estimator, while for $\delta > 1$ the randomized PPS estimator is more efficient than the ratio estimator.

Lemma 1. Let $\text{Eff}_{P/R}$ and ρ_z be defined by (14) and (16), respectively. If $V_x^2 > 0$, then $\text{Eff}_{P/R}$ is a monotonically increasing function of δ .

Proof. Write ρ_z from (16) as a weighted mean of the (negative) X_i

$$\rho_z = -u(\delta) = -\sum_{i \in U} w_i X_i,$$

where

$$w_i = \frac{X_i^{\delta-1}}{\sum_{i \in U} X_i^{\delta-1}} \quad [\text{Note that } \mu_x = u(2)].$$

Let $X_i > X_j$ ($i \neq j$), and define $h(\delta)$ as $w_i/w_j = (X_i/X_j)^{\delta-1}$. Since $h(\delta)$ is increasing in δ , the weight of the larger X_i is increasing compared to that of X_j when δ is increasing. Hence, $u(\delta)$ is increasing and ρ_z is decreasing in δ . It suffices therefore to show that $Eff_{P/R}$ is decreasing in ρ_z . Writing (14) as

$$Eff_{P/R} = \frac{-(N-n)}{\rho_z^{-1} + (n-1)},$$

it is seen that $Eff_{P/R}$ is decreasing in ρ_z indeed. This concludes the proof.

3.4 An alternative structure among the disturbances

Finally, suppose the variance of the disturbances in (15) is of the form:

$$\text{var}(\varepsilon_i) = c_1 X_i + c_2 X_i^2 \quad (0 < c_1, c_2 \leq 1).$$

See Kott (1988). For this case we obtain in analogy with (16)

$$\rho_z = -\sum_{i \in U} \omega_i X_i,$$

where

$$\omega_i = \frac{1 + \phi X_i}{\sum_{i \in U} (1 + \phi X_i)}, \text{ and } \phi = c_2 / c_1$$

when $\phi = 0$, $\rho_z = -1/N$. Hence, when $c_2 = 0$, PPS sampling is only as efficient as the ordinary ratio estimator from SRS sampling. Along the same lines as the proof of Lemma 1, it can be shown that ρ_z is decreasing in ϕ while $Eff_{P/R}$ is increasing in ϕ . Hence, for this case the randomized PPS estimator is always more efficient than the ratio estimator when c_2 is positive.

4. An application to the Producer Price Index

The Producer Price Index (PPI) in The Netherlands is based on about 2,500 commodity price indexes organized by type of product. The price index for a specific commodity can be written as

$$Y = \sum_{i \in U} X_i Z_i,$$

where Z_i is the price change for that commodity of establishment i relative to the basic period while X_i is the relative sales of that commodity by establishment i in the basic period (recall $\sum X_i = 1$).

In the example given here, we examine the price changes of 70 establishments for the commodity *Basic Metal* in December of 2005 relative to December of 2004; see Table 1. We compare the variance of the ratio estimator from an SRS sample with the variance of the HT estimator from a randomized PPS sample when $n = 9$. Applying (12) to these data gives $\text{var}(\hat{Y}_R) = 101$. If the sample had been drawn with replacement the variance would have been 116. Applying (3) and (9) for a randomized PPS sample gives $\text{var}(\hat{Y}_{PPS,\gamma}) = 29.9$. This outcome takes γ into account and lies close to the result $V_{PPS}^{(sim)} = 29.2$ from a simulation experiment consisting of 80,000 randomized PPS samples of size $n = 9$ from the set of 70 establishments. Hence, $Eff_{P/R} = 3.5$. Because formula (12) for $\text{var}(\hat{Y}_R)$ is only asymptotically unbiased, we also carried out simulations evaluating the mean square error (MSE) and the bias of \hat{Y}_R resulting in $\text{MSE}_R^{(sim)} = 108$ and a relatively small bias of 0.7. This confirms the conjecture that (12) gives an underestimation of the true variance; see Cochran (1977). Hence, for moderate samples the true value of $Eff_{P/R}$ might be somewhat higher than (14) suggests.

Furthermore, it is noteworthy that the simpler formula (10) for ρ_z in combination with (3) gives almost the same result $\text{var}(\hat{Y}_{PPS}) = 30.7$ even though $N = 70$ is not very large. The with replacement PPS variance would have been 43.8. Hence, the variance reduction for randomized PPS sampling is more than 30% even though the sampling fraction n/N is much smaller. According to (18), formula (12) with $n_{SRS} = 26$ gives about the same outcome as (3) with $n_{PPS} = 9$; note: $\rho_z = -0.042$. Hence, the sample sizes differ by a factor 2.9, which is more or less in line with the factor $(1 + CV_x^2) = 3.1$ from subsection 3.2. This should not be surprising because the price changes and their variability hardly depend on the sizes of the company. Fitting a double log regression

$$\ln(Z_i - Y)^2 = \alpha + \beta \ln X_i + v_i \quad (19)$$

results in the estimate $\hat{\beta} = 0.07$ for the data in Table 1; units with $Z_i = Y$ should be omitted in the regression. The estimate $\hat{\beta} = 0.07$ corresponds with $\hat{\delta} = 2.07$ for the disturbances in (15) which explains the superiority of randomized PPS sampling for this type of data. Also for other commodities $\hat{\delta}$ often was about 2; see Enthoven (2007).

Table 1
Price changes (Z_i) and sizes (X_i) of 70 establishments

i	price change	size	i	price change	size
1	-18.4%	0.0608	36	34.8%	0.0427
2	-16.0%	0.0784	37	13.1%	0.0121
3	3.3%	0.0762	38	31.7%	0.0351
4	12.5%	0.0100	39	-24.8%	0.0074
5	0.0%	0.0029	40	55.3%	0.0009
6	8.3%	0.0006	41	40.5%	0.0066
7	-39.0%	0.0182	42	34.6%	0.0022
8	-25.1%	0.0020	43	1.7%	0.0001
9	1.1%	0.0040	44	0.0%	0.0039
10	4.4%	0.0066	45	3.9%	0.0304
11	-4.9%	0.0039	46	25.4%	0.0209
12	-8.9%	0.0070	47	25.6%	0.0062
13	-7.0%	0.0148	48	0.0%	0.0033
14	-15.0%	0.0108	49	-0.3%	0.0019
15	-10.7%	0.0087	50	66.6%	0.0346
16	-9.0%	0.1079	51	0.0%	0.0039
17	-11.3%	0.0247	52	-2.9%	0.0007
18	10.6%	0.0024	53	15.8%	0.0011
19	-23.2%	0.0001	54	0.0%	0.0026
20	-25.4%	0.0001	55	0.0%	0.0018
21	-80.7%	0.0002	56	11.6%	0.0057
22	13.4%	0.0005	57	0.0%	0.0042
23	-42.5%	0.0010	58	0.0%	0.0236
24	-34.8%	0.0014	59	-1.5%	0.0015
25	-30.0%	0.0126	60	0.0%	0.0003
26	8.0%	0.0530	61	11.7%	0.0067
27	0.0%	0.0208	62	0.0%	0.0012
28	2.1%	0.0119	63	0.8%	0.0040
29	11.3%	0.0208	64	2.0%	0.0009
30	0.7%	0.0322	65	2.3%	0.0018
31	9.5%	0.0447	66	4.7%	0.0026
32	11.5%	0.0018	67	0.9%	0.0064
33	5.8%	0.0174	68	-1.0%	0.0309
34	-6.9%	0.0197	69	-0.5%	0.0005
35	0.0%	0.0124	70	0.0%	0.0006

We conclude this section with a small example showing that randomized PPS is not *always* better than the ratio estimator. Although the data in Table 2 for a population of five units are artificial, a data pattern like this may occur in financial branches where very small financial companies may grow very fast with respect to certain financial variables. This high variability among growth rates of small companies results in a low value for δ . For an SRS sample with $n = 2$ from the five units in Table 2 the variance of the ratio estimator is 211 according to (12); simulations give $\text{MSE}_R^{(sim)} = 323$. This is much less than the variance of 557 found in a simulation consisting of 80,000 randomized PPS samples of size $n = 2$. Formula (3) in combination with (9) gives the same outcome: 557. This would also be the correct variance had sample been drawn according to Brewer

(1963a) or Durbin (1967). Formula (11), based on (10), gives a slightly different value, 556.

Regression (19) with the data from Table 2 yields $\hat{\beta} = -3.0$, and hence $\hat{\delta} = -1.0$. In line with the findings of subsection 3.3 this low value $\hat{\delta} = -1.0$ explains why \hat{Y}_{PPS} is less efficient than \hat{Y}_R in this example. Moreover, the ordinary direct estimator $N\bar{y}_s$ from an SRS sample has a variance of 356, which is even smaller here than the variance in randomized PPS sampling; \bar{y}_s being the sample mean of the Y_i . Hence, for this type of data, the ratio estimator is the best option. Recall that the ratio estimator has a smaller variance than $N\bar{y}_s$ when $b > Y/2X$ where b is the slope of a regression from Y_i on X_i and a constant ($i = 1, \dots, N$); see Knottnerus (2003, page 117). So the data $Y_i (= X_i Z_i)$ in Table 2 certainly do not exhibit a flat trend.

Table 2
Growth rates of assets (Z_i) and sizes (X_i) of 5 establishments

i	growth rate	size
1	200%	0.0455
2	33%	0.1364
3	75%	0.1818
4	33%	0.2727
5	62%	0.3636

5. Summary

This paper compares the variance of the HT estimator \hat{Y}_{PPS} from a randomized PPS sample with the variance of the classical ratio estimator \hat{Y}_R from an SRS sample of the same size. In this comparison the sampling autocorrelation coefficient ρ_z plays an important role.

When the data pattern of the variables x and z ($= y/x$) is such that $\rho_z < -1/(N-1)$, it can be shown under mild conditions that \hat{Y}_{PPS} is more efficient than \hat{Y}_R for sufficiently large n and N , provided that X_i and Z_i are uncorrelated. Under model (15) with $E(\varepsilon_i^2) = \sigma^2 X_i^\delta$ it holds that $\rho_z < -1/(N-1)$ when $\delta > 1$. Hence, for this type of data \hat{Y}_{PPS} is to be preferred. Moreover, it emerges from (14) and (16) that for $\delta = 2$ the relative efficiency of PPS sampling compared to that of the ratio estimator is increasing when CV_x is increasing. In addition, \hat{Y}_R is to be preferred when the data correspond to a model with $\delta < 1$. These findings are confirmed empirically with a simulation study using two different data sets. When model (15) is not applicable, the relative efficiency of \hat{Y}_{PPS} is given by (14) provided n is large and N is relatively larger. In practice the unknown ρ_z in (14) is replaced by $\hat{\rho}_{z,9}$. The fact that $n \ll N$ does not necessarily mean that the factor $(n-1)\rho_z$ in (3) is always negligible.

Acknowledgements

The views expressed in the article are those of the author and do not necessarily reflect the policy of Statistics Netherlands. The author would like to thank Peter-Paul de Wolf, Sander Scholtus, the Associate Editor and two anonymous referees for their helpful suggestions and corrections.

Appendix A

A counterexample

Equations (5) and (7) cannot always be used for randomized PPS sampling when n and N are of the same

order while X_i and Z_i are correlated. To see that, consider a population U consisting of two groups U_1 and U_2 with means \bar{Y}_1 and \bar{Y}_2 , respectively. Both stratum sizes are $N/2$. Let s be a randomized PPS sample of size $n = 3N/4$ from the whole population U . Let the X_i be such that

$$\pi_i = nX_i = \begin{cases} 1 & \text{if } i \in U_1 \\ 0.5 & \text{if } i \in U_2. \end{cases}$$

Obviously, group 1 does not contribute to the variance. The selected units in s from U_2 constitute an ordinary SRS sample of size $N/4$. Hence, for randomized PPS sampling the correct variance formula in this example is

$$\text{var}(\hat{Y}_{PPS}) = \left(\frac{N}{2}\right)^2 \left(1 - \frac{1}{2}\right) \frac{S_{y2}^2}{N/4} = \frac{NS_{y2}^2}{2},$$

and

$$S_{y2}^2 = \frac{2}{N-2} \sum_{i \in U_2} (Y_{2i} - \bar{Y}_2)^2.$$

However, approximation (11) gives an entirely different, larger outcome unless $\bar{Y}_1 = 2\bar{Y}_2$.

References

- Brewer, K.R.W. (1963a). A model of systematic sampling with unequal probabilities. *Australian Journal of Statistics*, 5, 5-13.
- Brewer, K.R.W. (1963b). Ratio estimation and finite population: Some results deductible from the assumption of an underlying stochastic process. *Australian Journal of Statistics*, 5, 93-105.
- Brewer, K.R.W., and Donadio, M.E. (2003). The high entropy variance of the Horvitz-Thompson estimator. *Survey Methodology*, 29, 189-196.
- Cochran, W.G. (1977). *Sampling Techniques*. New York: John Wiley & Sons, Inc.
- Connor, W.S. (1966). An exact formula for the probability that two specified sampling units will occur in a sample drawn with unequal probabilities and without replacement. *Journal of the American Statistical Association*, 61, 384-390.
- Cumberland, W.G., and Royall, R.M. (1981). Prediction models and unequal probability sampling. *Journal of the Royal Statistical Society*, B, 43, 353-367.
- Deville, J.-C. (1999). Variance estimation for complex statistics and estimators: Linearization and residual techniques. *Survey Methodology*, 25, 193-203.
- Durbin, J. (1967). Design of multi-stage surveys for the estimation of sampling errors. *Applied Statistics*, 16, 152-164.
- Enthoven, L. (2007). *Cohort calculations* (in Dutch). Report MIC-2007-21, Statistics Netherlands, Voorburg.
- Foreman, E.K., and Brewer, K.R.W. (1971). The efficient use of supplementary information in standard sampling procedures. *Journal of the Royal Statistical Society*, B, 33, 391-400.

- Gabler, S. (1984). On unequal probability sampling: Sufficient conditions for the superiority of sampling without replacement. *Biometrika*, 71, 171-175.
- Hansen, M.H., and Hurwitz, W.N. (1943). On the theory of sampling from finite populations. *Annals of Mathematical Statistics*, 14, 333-362.
- Hartley, H.O., and Rao, J.N.K. (1962). Sampling with unequal probabilities and without replacement. *Annals of Mathematical Statistics*, 33, 350-374.
- Hidiroglou, M.A., and Gray, G.B. (1980). Construction of joint probability of selection for systematic P.P.S. sampling. *Applied Statistics*, 29, 107-112.
- Horvitz, D.G., and Thompson, D.J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47, 663-685.
- Knottnerus, P. (2003). *Sample Survey Theory: Some Pythagorean Perspectives*. New York: Springer-Verlag.
- Kott, P.S. (1988). Model-based finite population correction for the Horvitz-Thompson estimator. *Biometrika*, 75, 797-799.
- Kott, P.S. (2005). A note on the Hartley-Rao variance estimator. *Journal of Official Statistics*, 21, 433-439.
- Qualité, L. (2008). A comparison of conditional Poisson sampling versus unequal probability sampling with replacement. *Journal of Statistical Planning and Inference*, 138, 1428-1432.
- Rosén, B. (1997). On sampling with probability proportional to size. *Journal of Statistical Planning and Inference*, 62, 159-191.
- Särndal, C.-E., Swensson, B. and Wretman, J.H. (1992). *Model Assisted Survey Sampling*. New York: Springer-Verlag.
- Sen, A.R. (1953). On the estimate of the variance in sampling with varying probabilities. *Journal of the Indian Society of Agricultural Statistics*, 5, 119-127.
- Stehman, S.V., and Overton, W.S. (1994). Comparison of variance estimators of the Horvitz-Thompson estimator for randomized variable probability systematic sampling. *Journal of the American Statistical Association*, 89, 30-43.
- Thompson, M.E., and Wu, C. (2008). Simulation-based randomized systematic PPS sampling under substitution of units. *Survey Methodology*, 34, 3-10.
- Yates, F., and Grundy, P.M. (1953). Selection without replacement from within strata with probability proportional to size. *Journal of the Royal Statistical Society, B*, 15, 253-261.

The use of estimating equations to perform a calibration on complex parameters

Éric Lesage¹

Abstract

In the calibration method proposed by Deville and Särndal (1992), the calibration equations take only exact estimates of auxiliary variable totals into account. This article examines other parameters besides totals for calibration. Parameters that are considered complex include the ratio, median or variance of auxiliary variables.

Key Words: Calibration; Complex parameter; Estimating equation; Calibration weight.

1. Introduction

In survey statistics, two main approaches are used in the estimation phase: “model-assisted” estimators (such as the regression estimator or the ratio estimator) and calibration estimators (such as the raking ratio), proposed by Deville and Särndal (1992). The two approaches are somewhat similar, as shown by the regression estimator, which is the same as the calibration estimator with the χ^2 distance (“linear” calibration method).

The purpose of this article is to expand the family of calibration estimators. With the current method, calibration can be performed on totals. The idea is to be able to take into account the calibration constraints of complex parameters or statistics such as a ratio, a median or a geometric mean. The reason for doing this is that auxiliary information may consist of a complex statistic rather than totals. For example, a ratio relative to the total population might be known, but not the total in the numerator or denominator.

The issue of complex parameters in calibrations has been discussed in the literature. Särndal (2007) reviewed a number of them, in particular the work of Harms and Duchesne (2006) on the calibration estimation of quantiles, and the work of Krapavickaitė and Plikusas (2005) on calibration estimators of certain functions of totals.

The originality of the approach in this article is that it reduces calibration on a complex parameter to calibration on a total for a new *ad hoc* auxiliary variable. The advantage of this approach is that current calibration tools can be used and that there is no need to solve a complex optimization program.

In section 2 of the article, we review how the calibration method works, define calibration on complex parameters and describe simple cases in which calibration on a complex parameter can be reduced to calibration on a total. In section 3, we focus on parameters that can be defined as a solution to an estimating equation (Godambe and Thompson 1986). We introduce the concept of calibration

on a complex parameter defined by an estimating equation and show that the resulting calibration equation can be replaced with an equation for calibration on a total.

2. A complex parameter defined as a function of totals

2.1 Review of calibration on totals

Let U be a finite population of size N . The statistical units of the population are indexed by a label k , where $k \in \{1, \dots, N\}$. A sample s is selected using sample plan $p(s)$. Its size is denoted n and may be random. Let π_k be the probability that k is included in sample s , and let $d_k = 1 / \pi_k$ be its sampling weight.

For any variable z that takes the values z_k for the units in U indexed by k , the sum $t_z = \sum_{k \in U} z_k$ is referred to as the total of z over U .

Let $y^{(1)}, \dots, y^{(Q)}$ be Q variables of interest, whose values are known only for sample s , and let θ_y be the parameter of interest that is a function of the totals $t_{y^{(1)}}, \dots, t_{y^{(Q)}}$:

$$\theta_y = f(t_{y^{(1)}}, \dots, t_{y^{(Q)}}).$$

The estimator of θ_y is

$$\hat{\theta}_{y,\pi} = f(\hat{t}_{y^{(1)},\pi}, \dots, \hat{t}_{y^{(Q)},\pi}).$$

It is simply the function $f(\cdot, \dots, \cdot)$ with totals $t_{y^{(q)}}$ replaced by their Horvitz-Thompson estimator $\hat{t}_{y^{(q)},\pi} = \sum_{k \in s} d_k y_k^{(q)}$ (Särndal, Swensson and Wretman 1992). This estimator can be described as a substitution estimator.

Let $x^{(1)}, \dots, x^{(P)}$ be P auxiliary variables known on s , and let $t_{x^{(1)}}, \dots, t_{x^{(P)}}$ be the totals on U for those auxiliary variables, also known. For an individual k , the vector of values taken by the auxiliary variables on k is denoted $\mathbf{x}'_k = (x_k^{(1)}, \dots, x_k^{(P)})$.

The calibration estimator of θ_y is

1. Éric Lesage, CREST(ENSAI) and IRMAR(UEB), Ker Lann Campus, F-35172 BRUZ, France. E-mail: eric.lesage@ensai.fr.

$$\hat{\theta}_{y, \text{CAL}} = f(\hat{t}_{y^{(1)}, \text{CAL}}, \dots, \hat{t}_{y^{(Q)}, \text{CAL}})$$

with $\hat{t}_{y^{(q)}, \text{CAL}} = \sum_{k \in s} w_k y_k^{(q)}$, and a series of weights $\{w_k\}_{k \in s}$, known as calibration weights (which should be denoted $w_k(s)$, since they depend on the sampling), obtained by solving the following optimization program:

$$\min_{\{w_k\}_{k \in s}} \sum_{k \in s} d(w_k, d_k)$$

under constraints

$$\begin{cases} \hat{t}_{x^{(1)}, \text{CAL}} = t_{x^{(1)}} \\ \dots \\ \hat{t}_{x^{(P)}, \text{CAL}} = t_{x^{(P)}} \end{cases}$$

$d(\cdot, \cdot)$ is a pseudo-distance, i.e., a function that measures the difference between the calibration weight and the sampling weight (unlike a difference, a pseudo-distance is not necessarily symmetrical on its two arguments). The program is solved with a Lagrangian. When the distance used is the χ^2 distance (i.e., $d(w_k, d_k) = (1/2)(w_k - d_k)^2/d_k$), the solution is $w_k = d_k(1 + \mathbf{x}'_k \lambda)$ (where λ is a P -vector of Lagrange multipliers).

2.2 Calibration on a complex parameter η_x

Definition 1: Let $x^{(1)}, \dots, x^{(P)}$ be P auxiliary variables known on s , and let $\eta_x = g(t_{x^{(1)}}, \dots, t_{x^{(P)}})$ be a complex parameter, a function of the totals of those auxiliary variables, also known.

In the case of calibration on the complex parameter η_x , the calibration weights are obtained by solving the following optimization program:

$$\min_{\{w_k\}_{k \in s}} \sum_{k \in s} d(w_k, d_k)$$

under constraints

$$\hat{\eta}_{x, \text{CAL}} = g(\hat{t}_{x^{(1)}, \text{CAL}}, \dots, \hat{t}_{x^{(P)}, \text{CAL}}) = \eta_x.$$

The totals $t_{x^{(q)}}$ do not have to be known, but the complex parameter η_x does.

Consider the example of the ratio

$$R_x = \frac{t_{x^{(1)}}}{t_{x^{(2)}}} = \frac{\sum_{k \in U} x_k^{(1)}}{\sum_{k \in U} x_k^{(2)}}.$$

The calibration estimator of R_x is of the form

$$\hat{R}_{x, \text{CAL}} = \frac{\sum_{k \in s} w_k x_k^{(1)}}{\sum_{k \in s} w_k x_k^{(2)}}.$$

The calibration equation in the case of calibration on a ratio is

$$\hat{R}_{x, \text{CAL}} = \frac{\sum_{k \in s} w_k x_k^{(1)}}{\sum_{k \in s} w_k x_k^{(2)}} = R_x$$

R_x is known auxiliary information, as the total of the auxiliary variables usually is. This scenario may occur when we have proportions that are well known and stable over time, for example, but the specific totals in the numerator and denominator are not known.

We described the case of calibration on a single complex parameter, but it is clearly a simple matter to calibrate on more than one complex parameter. In that case, there are as many constraints as calibration parameters.

2.3 Simple cases where calibration on a complex parameter can be reduced to calibration on a total

It is not easy to determine from the outset whether an equation for calibration on a complex parameter can be written in the form of an equation for calibration on a total. In other words, it is not always a trivial matter to find a “new” auxiliary variable z , associated with the complex parameter, on whose total we can calibrate.

For example, that is quite straightforward for all moments of an auxiliary variable x (it is assumed that under the sampling plan, the population size N can be estimated exactly). If $\mu_{x^m} = N^{-1} \sum_{k \in U} x_k^m$ is auxiliary information, we can simply take $z_k = x_k^m / N$ and calibrate on μ_{x^m} : $\sum_{k \in s} w_k x_k^m / N = \mu_{x^m}$.

If we want to calibrate on the variance and the mean of variable x with μ_x and σ_x^2 as auxiliary information, we can use the two new auxiliary variables

$$z_k^{(1)} = \frac{x_k}{N}$$

and

$$z_k^{(2)} = \frac{(x_k - \mu_x)^2}{N}.$$

On the other hand, if we do not know μ_x , but we have σ_x^2 in the auxiliary information and we want to calibrate on that variance, things become more complicated. We can see this if we write the substitution estimator of σ_x^2 (where the sampling plan allows the population size N to be estimated exactly):

$$\hat{\sigma}_{x, \text{CAL}}^2 = \frac{1}{N} \sum_{k \in s} w_k \left(x_k - \left(\frac{\sum_{l \in s} w_l x_l}{N} \right) \right)^2.$$

Finding a new auxiliary variable z is not straightforward, since the initial calibration equation is not linear relative to the weight vector. We will return to the variance case in section 3.3 below.

Ratio example

Proposition 1: Calibration on a ratio is equivalent to calibration on the total of the new auxiliary variable: $z_k = x_k^{(1)} - R_x x_k^{(2)}$.

The calibration equation is written

$$\hat{t}_{z,CAL} = t_z = 0.$$

Proof:

$$\hat{t}_{z,CAL} = t_z$$

$$\Leftrightarrow \sum_{k \in S} w_k (x_k^{(1)} - R_x x_k^{(2)}) = \sum_{k \in U} (x_k^{(1)} - R_x x_k^{(2)})$$

$$\Leftrightarrow \hat{t}_{x^{(1)},CAL} - R_x \hat{t}_{x^{(2)},CAL} = t_{x^{(1)},CAL} - R_x t_{x^{(2)},CAL} = 0$$

$$\Leftrightarrow \frac{\hat{t}_{x^{(1)},CAL}}{\hat{t}_{x^{(2)},CAL}} = R_x$$

$$\text{i.e., } \hat{R}_{x,CAL} = R_x.$$

Function of a ratio of linear combinations of totals

Let η_x be a complex parameter that is a bijective function of a ratio of linear combinations of totals:

$$\eta_x = h \left(\frac{\alpha' \cdot \mathbf{t}_x}{\beta' \cdot \mathbf{t}_x} \right) \quad (1)$$

with $\alpha' = (\alpha_1, \dots, \alpha_p)$ and $\beta' = (\beta_1, \dots, \beta_p)$ being vectors of real coefficients of size P , and $\mathbf{t}'_x = (t_{x^{(1)}}, \dots, t_{x^{(p)}})$.

Proposition 2: Performing a calibration on complex parameter η_x defined by function (1) is equivalent to calibrating on the total of the new auxiliary variable:

$$z_k = (\alpha' - h^{-1}(\eta_x) \beta') \cdot \mathbf{x}_k$$

with calibration equation

$$\hat{t}_{z,CAL} = \sum_{k \in S} w_k z_k = t_z = 0.$$

Proof:

$$\hat{\eta}_{x,CAL} = \eta_x \Leftrightarrow h \left(\frac{\alpha' \cdot \hat{\mathbf{t}}_{x,CAL}}{\beta' \cdot \hat{\mathbf{t}}_{x,CAL}} \right) = \eta_x$$

$$\Leftrightarrow \frac{\alpha' \cdot \hat{\mathbf{t}}_{x,CAL}}{\beta' \cdot \hat{\mathbf{t}}_{x,CAL}} = h^{-1}(\eta_x)$$

$$\Leftrightarrow (\alpha' - h^{-1}(\eta_x) \beta') \cdot \hat{\mathbf{t}}_{x,CAL} = 0$$

$$\Leftrightarrow \sum_{k \in S} w_k (\alpha' - h^{-1}(\eta_x) \beta') \cdot \mathbf{x}_k = 0.$$

Consider the example of the geometric mean:

$$\mu_{Geo,x} = \left(\prod_{k \in U} x_k \right)^{1/N}.$$

This expression can be rewritten as

$$\mu_{Geo,x} = \exp \left(\frac{\sum_{k \in U} \ln(x_k)}{\sum_{k \in U} 1} \right).$$

We denote $\mathbf{x}'_k = (x_k^{(1)}, x_k^{(2)}) = (\ln(x_k), 1)$, $\alpha' = (1, 0)$, $\beta' = (0, 1)$ and $h^{-1}(u) = \exp^{-1}(u) = \ln(u)$.

Hence, the new auxiliary variable is

$$z_k = \ln(x_k) - \ln(\mu_{Geo,x}) \cdot 1.$$

We will see later in the article that the estimating equations method provides another approach to displaying the new auxiliary variable(s) \mathbf{z} .

3. Parameter defined by an estimating equation

3.1 Estimating with an estimating equation

Certain parameters θ_y are defined, or can be defined, as the solution to an implicit function known as the *estimating equation on U* (Godambe and Thompson 1986), i.e.:

$$\sum_{k \in U} \Phi(\theta_y, \mathbf{y}_k) = 0$$

with $\mathbf{y}'_k = (y_k^{(1)}, \dots, y_k^{(Q)})$ being the vector of values taken by the variables of interest for individual k .

In this context, an estimator of θ_y is defined for sample s , denoted $\hat{\theta}_{y,ee,\pi}$, which is the solution of the *estimating equation on s* (see in particular Hidiroglou, Rao and Yung 2002):

$$\sum_{k \in S} d_k \Phi(\hat{\theta}_{y,ee,\pi}, \mathbf{y}_k) = 0.$$

Table 1
Examples of parameters defined by estimating equations on U

Parameter	$\Phi(\theta_y, \mathbf{y}_k)$	Estimating equation on U
mean μ	$(\mathbf{y}_k - \mu)$	$\sum_{k \in U} (y_k - \mu) = 0$
ratio $R = \mu_1 / \mu_2$	$(y_k^{(1)} - R y_k^{(2)})$	$\sum_{k \in U} (y_k^{(1)} - R y_k^{(2)}) = 0$
median m	$(1_{y_k \leq m} - 1/2)$	$\sum_{k \in U} (1_{y_k \leq m} - 1/2) = 0$

Consider also the example of the coefficient of a logistic regression. Let $y^{(1)}$ be a dichotomous variable that takes the values 0 and 1 on U , and let $y^{(2)}$ be a quantitative variable. The value $y_k^{(1)}$ taken by $y^{(1)}$ for unit k is assumed to be an instance of the random variable $Y_k^{(1)}$, which has a Bernoulli distribution

$$\mathfrak{B} \left(1, p_k = \frac{1}{1 + \exp(-\beta_0 y_k^{(2)})} \right).$$

We have limited the number of parameters to one, but it would be just as simple to consider the multidimensional case. However, we should provide a definition of the estimating equations that take the case of the vector parameters into account.

The parameter of interest to us is the estimator of β_0 , denoted β , calculated on the finite population by the maximum likelihood method. The estimating equation of β on U will be the maximum likelihood equation. The log-likelihood in the case of Bernoulli variables is

$$L(\beta) = \sum_{k \in U} y_k^{(1)} \ln(p_k) + \sum_{k \in U} (1 - y_k^{(1)}) \ln(1 - p_k).$$

It is easy to derive the estimating equation of β on U :

$$\sum_{k \in U} y_k^{(2)} \left(y_k^{(1)} - \frac{1}{1 + \exp(-\beta y_k^{(2)})} \right) = 0.$$

The estimating equation on s which defines the estimator $\hat{\beta}_{ee,\pi}$ on the basis of the sampling weights is

$$\sum_{k \in s} d_k y_k^{(2)} \left(y_k^{(1)} - \frac{1}{1 + \exp(-\hat{\beta}_{ee,\pi} y_k^{(2)})} \right) = 0.$$

The estimating equation is not linear in the parameter; $\hat{\beta}_{ee,\pi}$ cannot be expressed as a simple function of the observations.

The logistic regression example is very interesting because it shows that we do not need to know $\hat{\beta}_{ee,\pi}$ to perform the calibration. We will see in the next subsection that we only need to know the generic term of the estimating equation on

$$U, \Phi(\beta, \mathbf{y}_k) = y_k^{(2)} \left(y_k^{(1)} - \frac{1}{1 + \exp(-\beta y_k^{(2)})} \right),$$

for all $k \in s$.

3.2 Calibration in the case of parameters defined by estimating equations

Let $\mathbf{x}'_k = (x_k^{(1)}, \dots, x_k^{(P)})$ be the vector of P known auxiliary variables on s , and let $\eta_{\mathbf{x}}$ be a complex parameter, also known, defined by the estimating equation

$$\sum_{k \in U} \Psi(\eta_{\mathbf{x}}, \mathbf{x}_k) = 0.$$

Definition 2: In the case of calibration on the complex parameter $\eta_{\mathbf{x}}$, the calibration weights are obtained by solving the following optimization program:

$$\min_{\{w_k\}_{k \in s}} \sum_{k \in s} d(w_k, d_k)$$

under constraints

$$\sum_{k \in s} w_k \Psi(\eta_{\mathbf{x}}, \mathbf{x}_k) = 0.$$

Proposition 3: Calibration on a complex parameter $\eta_{\mathbf{x}}$, defined by an estimating equation, is equivalent to a calibration on the total of the new auxiliary variable: $z_k = \Psi(\eta_{\mathbf{x}}, \mathbf{x}_k)$, with the calibration constraint $\sum_{k \in s} w_k z_k = 0$.

Definition 3: A calibration estimator of the parameter of interest θ_y , denoted $\hat{\theta}_{y,ee,CAL}$, is a solution to the estimating equation on s weighted by the calibration weights $\{w_k\}_{k \in s}$:

$$\sum_{k \in s} w_k \Phi(\hat{\theta}_{y,ee,CAL}, \mathbf{y}_k) = 0.$$

In most cases, the solution to the estimating equation is unique. The median is an example of a parameter for which there may be more than one solution. In this case, the infimum is often used as an estimator.

Proposition 4: If there is only one solution to the equation $\sum_{k \in s} w_k \Psi(\hat{\eta}_{\mathbf{x},ee,CAL}, \mathbf{x}_k) = 0$, then

$$\hat{\eta}_{\mathbf{x},ee,CAL} = \eta_{\mathbf{x}}.$$

Proof: $\eta_{\mathbf{x}}$ is a solution to the estimating equation that defines $\hat{\eta}_{\mathbf{x},ee,CAL}$. Since there is a unique solution, we have $\hat{\eta}_{\mathbf{x},ee,CAL} = \eta_{\mathbf{x}}$.

3.3 Calibration on a variance

In this section, we examine calibration on variance σ_x^2 , which is a more complicated complex parameter than those discussed above. We will show that when the variance is the only auxiliary information we have, we can perform an approximate calibration that produces calibration weights that have better properties than the sampling weights.

Back to the variance case. The mean μ_x and the variance σ_x^2 on U of auxiliary variable x can be defined by two estimating equations on U :

$$\begin{cases} \sum_{k \in U} (x_k - \mu_x) = 0 \end{cases} \quad (2)$$

$$\begin{cases} \sum_{k \in U} ((x_k - \mu_x)^2 - \sigma_x^2) = 0. \end{cases} \quad (3)$$

If we know the two parameters, calibrating on them is easy, since we merely have to calibrate on the totals of the two new auxiliary variables $z^{(1)} = x - \mu_x$ and $z^{(2)} = (x - \mu_x)^2 - \sigma_x^2$.

On the other hand, if we consider the textbook case where the mean μ_x is not known, the parameter σ_x^2 cannot be defined by a unique estimating equation. If we replace μ_x with its explicit definition

$$\mu_x = \frac{\sum_{l \in \ell'} x_l}{\sum_{l \in \ell'} 1}$$

in equation (3), we obtain the equation

$$\sum_{k \in U} \left(\left(x_k - \frac{\sum_{l \in U} x_l}{\sum_{j \in U} 1} \right)^2 - \sigma_x^2 \right) = 0,$$

which cannot be written in the form of an estimating equation: $\sum_{k \in U} \Psi(\sigma_x^2, x_k) = 0$.

μ_x thus becomes a nuisance parameter (Binder 1991). To overcome this difficulty, we can replace it in equation (3) with its substitution estimator: $\hat{\mu}_{x,\pi} = \hat{t}_{x,\pi} / \hat{N}_\pi$, with $\hat{N}_\pi = \sum_{k \in s} d_k 1$ being the Horvitz-Thompson estimator of the size of population U . This leads to the “approximate” calibration equation

$$\sum_{k \in s} w_k \left(\left(x_k - \frac{\hat{t}_{x,\pi}}{\hat{N}_\pi} \right)^2 - \sigma_x^2 \right) = 0. \quad (4)$$

Proposition 5: With estimating equation (4), calibration on the variance is not perfect, and we have

$$\hat{\sigma}_{x,ee,CAL}^2 = \sigma_x^2 - \left(\frac{\hat{t}_{x,\pi}}{\hat{N}_\pi} - \frac{\hat{t}_{x,CAL}}{\hat{N}_{CAL}} \right)^2. \quad (5)$$

Proof:

- The “approximate” calibration equation is equation (4).
- The definition of the parameters’ calibration estimators:

$$\begin{cases} \sum_{k \in s} w_k (x_k - \hat{\mu}_{x,ee,CAL}) = 0 \\ \sum_{k \in s} w_k ((x_k - \hat{\mu}_{x,ee,CAL})^2 - \hat{\sigma}_{x,ee,CAL}^2) = 0. \end{cases}$$

This can be rewritten

$$\begin{cases} \hat{\mu}_{x,ee,CAL} = \frac{\sum_{k \in s} w_k x_k}{\sum_{k \in s} w_k} = \frac{\hat{t}_{x,CAL}}{\hat{N}_{CAL}} \\ \sum_{k \in s} w_k \left(\left(x_k - \frac{\hat{t}_{x,CAL}}{\hat{N}_{CAL}} \right)^2 - \hat{\sigma}_{x,ee,CAL}^2 \right) = 0. \end{cases}$$

- If we subtract the second estimating equation from the approximate calibration equation, we get

$$\sum_{k \in s} w_k \left(\left(x_k - \frac{\hat{t}_{x,\pi}}{\hat{N}_\pi} \right)^2 - \left(x_k - \frac{\hat{t}_{x,CAL}}{\hat{N}_{CAL}} \right)^2 - \sigma_x^2 + \hat{\sigma}_{x,ee,CAL}^2 \right) = 0.$$

Using the identity $a^2 - b^2 = (a - b)(a + b)$, we have

$$\sum_{k \in s} w_k \left(\left(\frac{\hat{t}_{x,CAL}}{\hat{N}_{CAL}} - \frac{\hat{t}_{x,\pi}}{\hat{N}_\pi} \right) \left(2x_k - \frac{\hat{t}_{x,\pi}}{\hat{N}_\pi} - \frac{\hat{t}_{x,CAL}}{\hat{N}_{CAL}} \right) - \hat{N}_{CAL} (\sigma_x^2 - \hat{\sigma}_{x,ee,CAL}^2) \right) = 0$$

$$\left(\frac{\hat{t}_{x,CAL}}{\hat{N}_{CAL}} - \frac{\hat{t}_{x,\pi}}{\hat{N}_\pi} \right) \sum_{k \in s} w_k \left(2x_k - \frac{\hat{t}_{x,\pi}}{\hat{N}_\pi} - \frac{\hat{t}_{x,CAL}}{\hat{N}_{CAL}} \right) - \hat{N}_{CAL} (\sigma_x^2 - \hat{\sigma}_{x,ee,CAL}^2) = 0$$

$$\left(\frac{\hat{t}_{x,CAL}}{\hat{N}_{CAL}} - \frac{\hat{t}_{x,\pi}}{\hat{N}_\pi} \right) \left(2\hat{t}_{x,CAL} - \frac{\hat{t}_{x,\pi}}{\hat{N}_\pi} \hat{N}_{CAL} - \hat{t}_{x,CAL} \right) - \hat{N}_{CAL} (\sigma_x^2 - \hat{\sigma}_{x,ee,CAL}^2) = 0$$

$$\hat{N}_{CAL} \left(\frac{\hat{t}_{x,CAL}}{\hat{N}_{CAL}} - \frac{\hat{t}_{x,\pi}}{\hat{N}_\pi} \right)^2 - \hat{N}_{CAL} (\sigma_x^2 - \hat{\sigma}_{x,ee,CAL}^2) = 0.$$

This is the same as the expression for $\hat{\sigma}_{x,ee,CAL}^2$ in equation (5).

This result is interesting because, without an exact calibration, we have a calibration estimator of σ_x^2 that is asymptotically more precise than the substitution estimator $\hat{\sigma}_{x,\pi}^2$. That is, if we resort to the asymptotic framework typically used in surveys and employ linearization of complex estimators (Deville 1999), we have

$$\hat{\sigma}_{x,\pi}^2 - \sigma_x^2 = O_p \left(\frac{1}{\sqrt{n}} \right)$$

and

$$(\hat{\sigma}_{x,ee,CAL}^2 - \sigma_x^2)^{1/2} = \left(\frac{\hat{t}_{x,\pi}}{\hat{N}_\pi} - \frac{\hat{t}_{x,CAL}}{\hat{N}_{CAL}} \right) = O_p \left(\frac{1}{\sqrt{n}} \right).$$

This yields

$$\hat{\sigma}_{x,ee,CAL}^2 - \sigma_x^2 = O_p \left(\frac{1}{n} \right).$$

4. Conclusion

In this article, we presented a simple method of performing a calibration in cases where the auxiliary information takes the form of a complex parameter. That method is based on the concept of the estimating equation. Its major advantage is that it can be used with current calibration software.

In future research, it would be interesting to determine the practical cases in which the use of complex parameters in the calibration improves the precision of the parameters of interest.

Acknowledgements

The author wishes to thank the journal's associate editor, the reviewers, Guillaume Chauvet, François Coquet and Jean-Claude Deville for their constructive comments on the preliminary versions of this paper.

References

- Binder, D.A. (1991). Use of estimating functions for interval estimation from complex surveys. *Proceedings of the survey research methods section*, American Statistical Association, 34-42.
- Deville, J.-C. (1999). Variance estimation for complex statistics and estimators: Linearization and residual techniques. *Survey Methodology*, 25, 193-203.
- Deville, J.-C., and Särndal, C.-E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, 87, 376-382.
- Godambe, V.P., and Thompson, M.E. (1986). Parameters of superpopulation and survey population: Their relationship and estimation. *International Statistical Review*, 54, 127-138.
- Harms, T., and Duchesne, P. (2006). On calibration estimation for quantiles. *Survey Methodology*, 32, 37-52.
- Hidiroglou, M., Rao, J.N.K. and Yung, W. (2002). Estimating equations for the analysis of survey data using poststratification information. *Sankhyā*, 64, 2, 364-378.
- Krapavickaitė, D., and Plikusas, A. (2005). Estimation of ratio in finite population. *Informatica*, 16, 347-364.
- Plikusas, A. (2006). Non-linear calibration. Recueil du Colloque sur les méthodes de sondage, Venspils, Latvia. Riga: Central Statistical Bureau of Latvia.
- Särndal, C.-E., Swensson, B. and Wretman, J. (1992). *Model Assisted Survey Sampling*. New-York: Springer-Verlag, 162-163.
- Särndal, C.-E. (2007). The calibration approach in survey theory and practice. *Survey Methodology*, 33, 99-119.

JOURNAL OF OFFICIAL STATISTICS

An International Review Published by Statistics Sweden

JOS is a scholarly quarterly that specializes in statistical methodology and applications. Survey methodology and other issues pertinent to the production of statistics at national offices and other statistical organizations are emphasized. All manuscripts are rigorously reviewed by independent referees and members of the Editorial Board.

Contents Volume 26, No. 4, 2010

Editorial	
Ingegerd Jansson and Boris Lorenc.....	i
Degrees of Freedom Approximations and Rules-of-Thumb	
Richard Valliant and Keith F. Rust	585
Combining Link-Tracing Sampling and Cluster Sampling to Estimate Totals and Means of Hidden Human Populations	
Martín H. Félix-Medina and Pedro E. Monjardin.....	603
Increasing Respondents' Use of Definitions in Web Surveys	
Andy Peytchev, Frederick G. Conrad, Mick P. Couper and Roger Tourangeau.....	633
A Framework for Cut-off Sampling in Business Survey Design	
Roberto Benedetti, Marco Bee and Giuseppe Espa.....	651
Statistical Model of the 2001 Czech Census for Interactive Presentation	
Jiří Grim, Jan Hora, Pavel Boček, Petr Somol and Pavel Pudil	673
An Optimal Multivariate Stratified Sampling Design Using Auxiliary Information: An Integer Solution Using Goal Programming Approach	
M.G.M. Khan, T. Maiti and M.J. Ahsan.....	695
Letter to the Editor	
Rainer Lenz.....	709
Book Reviews.....	711
Editorial Collaborators	717

All inquiries about submissions and subscriptions should be directed to jos@scb.se

JOURNAL OF OFFICIAL STATISTICS

An International Review Published by Statistics Sweden

JOS is a scholarly quarterly that specializes in statistical methodology and applications. Survey methodology and other issues pertinent to the production of statistics at national offices and other statistical organizations are emphasized. All manuscripts are rigorously reviewed by independent referees and members of the Editorial Board.

Contents

Volume 27, No. 1, 2011

The 2010 Morris Hansen Lecture Dealing with Survey Nonresponse in Data Collection, in Estimation Carl-Erik Särndal	1
Discussion	
J. Michael Brick	23
Roger Tourangeau	29
Breakoff and Unit Nonresponse Across Web Surveys Andy Peytchev	33
Assessing Mode Effects in a National Crime Victimization Survey using Structural Equation Models: Social Desirability Bias and Acquiescence Dirk Heerwegh and Geert Loosveldt	49
Designing Input Fields for Non-Narrative Open-Ended Responses in Web Surveys Mick P. Couper, Courtney Kennedy, Frederick G. Conrad and Roger Tourangeau	65
Using Register Data to Evaluate the Effects of Proxy Interviews in the Norwegian Labour Force Survey Ib Thomsen and Ole Villund	87
Linear Regression Influence Diagnostics for Unclustered Survey Data Jianzhu Li and Richard Valliant	99
Evaluating the Small-Sample Bias of the Delete-a-Group Jackknife for Model Analyses Phillip S. Kott and Steven T. Garren	121
Letter to the Editor James R. Knaub	135
Book Review	139
In Other Journals	149

All inquiries about submissions and subscriptions should be directed to jos@scb.se

CONTENTS

TABLE DES MATIÈRES

Volume 38, No. 4, December/décembre 2010

Abbas Khalili	
New estimation and feature selection methods in mixture-of-experts models	519
Iván A. Carrillo, Jiahua Chen and Changbao Wu	
The pseudo-GEE approach to the analysis of longitudinal surveys	540
Irène Gijbels, Marek Omelka and Dominik Sznajder	
Positive quadrant dependence tests for copulas	555
Hanfeng Chen, Jiahua Chen and Shun-Yi Chen	
Confidence intervals for the mean of a population containing many zero values under unequal-probability sampling	582
Mahmoud Torabi and Jon N.K. Rao	
Mean squared error estimators of small area means using survey weights	598
Zhiqiang Tan	
Nonparametric likelihood and doubly robust estimating equations for marginal and nested structural models	609
Xiaogang Duan, Jing Qin and Qihua Wang	
Optimal estimation in surrogate outcome regression problems	633
Jesse Frey	
Testing for equivalence of variances using Hartley's ratio	647
Yanyuan Ma and Guosheng Yin	
Semiparametric median residual life model and inference	665
Samiran Sinha	
An estimated-score approach for dealing with missing covariate data in matched case-control studies	680
Mahmoud Torabi and Rhonda J. Rosychuk	
Spatio-temporal modelling of disease mapping of rates	698
Guohua Yan, William J. Welch and Ruben H. Zamar	
Model-based linear clustering	716
Tingting Gou and Duncan Murdoch	
Simulation of extremes of diffusions	738

CONTENTS

TABLE DES MATIÈRES

Volume 39, No. 1, March/mars 2011

Baoying Yang, Gengsheng Qin and Jing Qin Empirical likelihood-based inferences for a low income proportion	1
Gengsheng Qin, Xiaoping Jin and Xiao-Hua Zhou Non-parametric interval estimation for the partial area under the ROC curve.....	17
Grace Y. Yi, Leilei Zeng and Richard J. Cook A robust pairwise likelihood method for incomplete longitudinal binary data arising in clusters	34
Konstantinos Kalogeropoulos, Petros Dellaportas and Gareth O. Roberts Likelihood-based inference for correlated diffusions.....	52
Chi-Chung Wen, Steve Y.H. Huang and Yau-Hung Chen Cox regression for current status data with mismeasured covariates	73
Jianqiang C. Wang and Mary C. Meyer Testing the monotonicity or convexity of a function using regression splines.....	89
Weixing Song and Juan Du A note on testing the regression functions via nonparametric smoothing	108
Hirokazu Yanagihara, Ken-Ichi Kamo and Tetsuji Tonda Second-order bias-corrected AIC in multivariate normal linear models under non-normality	126
Li-Chun Zhang and Nina Hagesæther A domain outlier robust design and smooth estimation approach.....	147
Steven N. MacEachern and Subharup Guha Parametric and semiparametric hypotheses in the linear model.....	165

GUIDELINES FOR MANUSCRIPTS

Before finalizing your text for submission, please examine a recent issue of *Survey Methodology* (Vol. 32, No. 2 and onward) as a guide and note particularly the points below. Articles must be submitted in machine-readable form, preferably in Word. A pdf or paper copy may be required for formulas and figures.

1. Layout

- 1.1 Documents should be typed entirely double spaced with margins of at least 1½ inches on all sides.
- 1.2 The documents should be divided into numbered sections with suitable verbal titles.
- 1.3 The name (fully spelled out) and address of each author should be given as a footnote on the first page of the manuscript.
- 1.4 Acknowledgements should appear at the end of the text.
- 1.5 Any appendix should be placed after the acknowledgements but before the list of references.

2. Abstract

The manuscript should begin with an abstract consisting of one paragraph followed by three to six key words. Avoid mathematical expressions in the abstract.

3. Style

- 3.1 Avoid footnotes, abbreviations, and acronyms.
- 3.2 Mathematical symbols will be italicized unless specified otherwise except for functional symbols such as “exp(·)” and “log(·)”, *etc.*
- 3.3 Short formulae should be left in the text but everything in the text should fit in single spacing. Long and important equations should be separated from the text and numbered consecutively with arabic numerals on the right if they are to be referred to later.
- 3.4 Write fractions in the text using a solidus.
- 3.5 Distinguish between ambiguous characters, (*e.g.*, w, ω; o, O, 0; l, 1).
- 3.6 Italics are used for emphasis.

4. Figures and Tables

- 4.1 All figures and tables should be numbered consecutively with arabic numerals, with titles that are as self explanatory as possible, at the bottom for figures and at the top for tables.

5. References

- 5.1 References in the text should be cited with authors' names and the date of publication. If part of a reference is cited, indicate after the reference, *e.g.*, Cochran (1977, page 164).
- 5.2 The list of references at the end of the manuscript should be arranged alphabetically and for the same author chronologically. Distinguish publications of the same author in the same year by attaching a, b, c to the year of publication. Journal titles should not be abbreviated. Follow the same format used in recent issues.

6. Short Notes

- 6.1 Documents submitted for the short notes section must have a maximum of 3,000 words.

DIRECTIVES CONCERNANT LA PRÉSENTATION DES TEXTES

Avant de finaliser votre texte pour le soumettre, prière d'examiner un numéro récent de *Techniques d'enquête* (à partir du vol. 32, N° 2) et de noter les points ci-dessous. Les articles doivent être soumis sous forme de fichiers de traitement de texte, préférablement Word. Une version pdf ou papier pourrait être requise pour les formules et graphiques.

1.	Présentation	1.1 Les textes doivent être écrits à double interligne partout et avec des marges d'au moins 1 1/2 pouce tout autour. 1.2 Les textes doivent être divisés en sections numérotées portant des titres appropriés. 1.3 Le nom (écrit au long) et l'adresse de chaque auteur doivent figurer dans une note au bas de la première page du texte. 1.4 Les remerciements doivent paraître à la fin du texte. 1.5 Toute annexe doit suivre les remerciements mais précéder la bibliographie.
2.	Résumé	Le texte doit commencer par un résumé composé d'un paragraphe suivi de trois à six mots clés. Éviter les expressions mathématiques dans le résumé.
3.	Rédaction	3.1 Éviter les notes au bas des pages, les abréviations et les sigles. 3.2 Les symboles mathématiques seront imprimés en italique à moins d'une indication contraire, sauf pour les symboles fonctionnels comme $\exp(\cdot)$ et $\log(\cdot)$ etc. 3.3 Les formules courtes doivent figurer dans le texte principal, mais tous les caractères dans le texte doivent correspondre à un espace simple. Les équations longues et importantes doivent être séparées du texte principal et numérotées en ordre consécutif par un chiffre arabe à la droite si l'auteur y fait référence plus loin. 3.4 Écrire les fractions dans le texte à l'aide d'une barre oblique. 3.5 Distinguer clairement les caractères ambigus (comme w, ω ; o, O, 0; l, I). 3.6 Les caractères italiques sont utilisés pour faire ressortir des mots.
4.	Figures et tableaux	4.1 Les figures et les tableaux doivent tous être numérotés en ordre consécutif avec des chiffres arabes et porter un titre aussi explicatif que possible (au bas des figures et en haut des tableaux).
5.	Bibliographie	5.1 Les références à d'autres travaux faites dans le texte doivent préciser le nom des auteurs et la date de publication. Si une partie d'un document est citée, indiquer laquelle après la référence. Exemple: Cochran (1977, page 164). 5.2 La bibliographie à la fin d'un texte doit être en ordre alphabétique et les titres d'un même auteur doivent être en ordre chronologique. Distinguer les publications d'un même auteur et d'une même année en ajoutant les lettres a, b, c, etc. à l'année de publication. Les titres de revues doivent être écrits au long. Suivre le modèle utilisé dans les numéros récents.
6.	Communications brèves	6.1 Les documents soumis pour la section des communications brèves doivent avoir au plus 3 000 mots.

CONTENTS

TABLE DES MATIÈRES

Volume 39, No. 1, March/mars 2011

Baoying Yang, Gengsheng Qin and Jing Qin Empirical likelihood-based inferences for a low income proportion	1
Gengsheng Qin, Xiaoping Jin and Xiao-Hua Zhou Non-parametric interval estimation for the partial area under the ROC curve.....	17
Grace Y. Yi, Leilei Zeng and Richard J. Cook A robust pairwise likelihood method for incomplete longitudinal binary data arising in clusters	34
Konstantinos Kalogeropoulos, Petros Dellaportas and Gareth O. Roberts Likelihood-based inference for correlated diffusions.....	52
Chi-Chung Wen, Steve Y. H. Huang and Yau-Hung Chen Cox regression for current status data with mismeasured covariates	73
Jianqiang C. Wang and Mary C. Meyer Testing the monotonicity or convexity of a function using regression splines.....	89
Weixing Song and Juan Du A note on testing the regression functions via nonparametric smoothing.....	108
Hirokazu Yanagihara, Ken-Ichi Kamo and Tetsuji Tonda Second-order bias-corrected AIC in multivariate normal linear models under non-normality	126
Li-Chun Zhang and Nina Hagesæther A domain outlier robust design and smooth estimation approach.....	147
Steven N. MacEachern and Subharup Guha Parametric and semiparametric hypotheses in the linear model.....	165

TABLE DES MATIÈRES

CONTENTS

Volume 38, No. 4, December/décembre 2010

Abbas Khalili	519
New estimation and feature selection methods in mixture-of-experts models	519
Iván A. Carrillo, Jiahua Chen and Changbao Wu	540
The pseudo-GEE approach to the analysis of longitudinal surveys	540
Irène Gijbels, Marek Omelka and Dominik Szajder	555
Positive quadrat dependence tests for copulas	555
Hanfeng Chen, Jiahua Chen and Shun-Yi Chen	582
Confidence intervals for the mean of a population containing many zero values under unequal-probability sampling	582
Mahmoud Torabi and Jon N.K. Rao	598
Mean squared error estimators of small area means using survey weights	598
Zhiqiang Tan	609
Nonparametric likelihood and doubly robust estimating equations for marginal and nested structural models	609
Xiaogang Duan, Jing Qin and Qihua Wang	633
Optimal estimation in surrogate outcome regression problems	633
Jesse Frey	647
Testing for equivalence of variances using Hartley's ratio	647
Yanyuan Ma and Guosheng Yin	665
Semiparametric median residual life model and inference	665
Samiran Sinha	680
An estimated-score approach for dealing with missing covariate data in matched case-control studies	680
Mahmoud Torabi and Rhonda J. Rosychuk	698
Spatio-temporal modelling of disease mapping of rates	698
Guohua Yan, William J. Welch and Ruben H. Zamar	716
Model-based linear clustering	716
Tingting Gou and Duncan Murdoch	738
Simulation of extremes of diffusions	738

JOURNAL OF OFFICIAL STATISTICS

An International Review Published by Statistics Sweden

JOS is a scholarly quarterly that specializes in statistical methodology and applications. Survey methodology and other issues pertinent to the production of statistics at national offices and other statistical organizations are emphasized. All manuscripts are rigorously reviewed by independent referees and members of the Editorial Board.

Contents
Volume 27, No. 1, 2011

The 2010 Morris Hansen Lecture Dealing with Survey Nonresponse in Data Collection, in Estimation
Carl-Erik Sæmndal 1

Discussion
J. Michael Brick 23
Roger Tourangeau 29

Breakoff and Unit Nonresponse Across Web Surveys
Andy Peytchev 33

Assessing Mode Effects in a National Crime Victimization Survey using Structural Equation
Models: Social Desirability Bias and Acquiescence
Dirk Heerwegh and Geert Loosveldt 49

Designing Input Fields for Non-Narrative Open-Ended Responses in Web Surveys
Mick P. Couper, Courtney Kennedy, Frederick G. Conrad and Roger Tourangeau 65

Using Register Data to Evaluate the Effects of Proxy Interviews in the Norwegian Labour Force Survey
Ib Thomsen and Ole Villund 87

Linear Regression Influence Diagnostics for Unclustered Survey Data
Jianzhu Li and Richard Valliant 99

Evaluating the Small-Sample Bias of the Delete-a-Group Jackknife for Model Analyses
Phillip S. Kott and Steven T. Gatten 121

Letter to the Editor
James R. Knaub 135

Book Review 139
In Other Journals 149

All inquiries about submissions and subscriptions should be directed to jos@scb.se

JOS is a scholarly quarterly that specializes in statistical methodology and applications. Survey methodology and other issues pertinent to the production of statistics at national offices and other statistical organizations are emphasized. All manuscripts are rigorously reviewed by independent referees and members of the Editorial Board.

JOURNAL OF OFFICIAL STATISTICS

An International Review Published by Statistics Sweden

Contents

Volume 26, No. 4, 2010

Editorial	Inggerd Jansson and Boris Lorenc.....	1
Degrees of Freedom Approximations and Rules-of-Thumb	Richard Valliant and Keith F. Rust	585
Combining Link-Tracing Sampling and Cluster Sampling to Estimate Totals and Means of Hidden Human Populations	Martin H. Félix-Medina and Pedro E. Monjardin.....	603
Increasing Respondents' Use of Definitions in Web Surveys	Andy Peytchev, Frederick G. Conrad, Mick P. Couper and Roger Tourangeau.....	633
A Framework for Cut-off Sampling in Business Survey Design	Roberto Benedetti, Marco Bee and Giuseppe Espa.....	651
Statistical Model of the 2001 Czech Census for Interactive Presentation	Jifi Grim, Jan Hora, Pavel Boček, Petr Somol and Pavel Pudil	673
An Optimal Multivariate Stratified Sampling Design Using Auxiliary Information:		
An Integer Solution Using Goal Programming Approach	M.G.M. Khan, T. Maiti and M.J. Ahsan.....	695
Letter to the Editor	Rainer Lenz.....	709
Book Reviews.....		711
Editorial Collaborators.....		717

All inquiries about submissions and subscriptions should be directed to jos@scb.se

$$\sigma_{x_{ee},CAL}^2 - \sigma_x^2 = O_p \left(\frac{1}{n} \right).$$

4. Conclusion

Cet article a présenté une méthode simple pour réaliser un calage dans le cas où l'information auxiliaire prend la forme d'un paramètre complexe. Cette méthode s'appuie sur la notion d'équation estimentante. Son gros avantage est qu'elle peut être mise en oeuvre avec les logiciels actuels de calage. Une piste de recherche intéressante serait d'étudier dans quels cas pratiques l'utilisation de paramètres complexes dans le calage améliore la précision des paramètres d'intérêt.

Remerciements

L'auteur remercie le rédacteur associé de la revue, les examinateurs, Guillaume Chauvet, François Coquet et Jean-Claude Deville pour leurs commentaires constructifs sur les versions provisoires de cette note.

- Binder, D.A. (1991). Use of estimating functions for interval estimation from complex surveys. *Proceedings of the survey research methods section*, American Statistical Association, 34-42.
- Deville, J.-C. (1999). Estimation de variance pour des statistiques et des estimateurs complexes : linéarisation et techniques des résidus. *Techniques d'enquête*, 25, 219-230.
- Deville, J.-C., et Särndal, C.-E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, 87, 376-382.
- Godambe, V.P., et Thompson, M.E. (1986). Parameters of superpopulation and survey population: Their relationship and estimation. *Revue Internationale de Statistique*, 54, 127-138.
- Harms, T., et Duchesne, P. (2006). De l'estimation des quantiles par calage. *Techniques d'enquête*, 32, 41-57.
- Hidiroglou, M., Rao, J.N.K. et Yung, W. (2002). Estimating equations for the analysis of survey data using poststratification information. *Sankhyā*, 64, 2, 364-378.
- Krapavickaitė, D., et Plikusas, A. (2005). Estimation of ratio in finite population. *Informatica*, 16, 347-364.
- Plikusas, A. (2006). Non-linear calibration. *Recueil du Colloque sur les méthodes de sondage*, Ventspils, Latvia. Riga : Central Statistical Bureau of Latvia.
- Särndal, C.-E., Swensson, B. et Wretman, J. (1992). *Model Assisted Survey Sampling*. New-York : Springer-Verlag, 162-163.
- Särndal, C.-E. (2007). La méthode de calage dans la théorie et la pratique des enquêtes. *Techniques d'enquête*, 33, 2, 113-135.

Bibliographie

$$(2) \quad \left\{ \sum_{k \in U} (x_k - \mu_x) = 0, \right. \\ \left. \sum_{k \in U} ((x_k - \mu_x)^2 - \sigma_x^2) = 0. \right. \quad (3)$$

Si on connaît ces deux paramètres, le calage sur ceux-ci est facile, puisqu'il suffit de caler sur les totaux des deux nouvelles variables auxiliaires $z^{(1)} = x - \mu_x$ et $z^{(2)} = (x - \mu_x)^2 - \sigma_x^2$.

Par contre, si on considère le cas d'école où la moyenne μ_x n'est pas connue, alors le paramètre σ_x^2 ne peut pas être défini par une équation estimante unique. En effet, si on remplace μ_x par sa définition explicite

$$\mu_x = \frac{\sum_{j \in U} x_j}{\sum_{j \in U} 1}$$

dans l'équation (3), on obtient l'équation

$$\sum_{k \in U} \left(x_k - \frac{\sum_{j \in U} x_j}{\sum_{j \in U} 1} \right)^2 - \sigma_x^2 = 0,$$

qui ne peut pas s'écrire sous la forme d'une équation estimante : $\sum_{k \in U} \Psi(\sigma_x^2, x_k) = 0$.

μ_x apparaît donc comme un paramètre de nuisance (Binder 1991). Pour contourner cette difficulté, on peut le remplacer dans l'équation (3) par son estimateur par substitution : $\hat{\mu}_{x,\pi} = \hat{t}_{x,\pi} / \hat{N}_\pi$, avec $\hat{N}_\pi = \sum_{k \in s} d_k$ l'estimateur Horvitz-Thompson de la taille de la population U .

On obtient alors comme équation de calage « approchée » :

$$(4) \quad \sum_{k \in s} w_k \left(x_k - \frac{\hat{t}_{x,\pi}}{\hat{N}_\pi} \right)^2 - \sigma_x^2 = 0.$$

Proposition 5 : Avec l'équation estimante (4), le calage sur la variance n'est pas parfait et nous avons :

$$(5) \quad \hat{\sigma}_{x,ee,CAL}^2 = \sigma_x^2 - \left(\frac{\hat{t}_{x,\pi}}{\hat{t}_{x,CAL}} - \frac{\hat{N}_\pi}{\hat{N}_{CAL}} \right)^2.$$

Démonstration :

- L'équation de calage « approchée » est l'équation (4).
- La définition des estimateurs par calage des paramètres :

$$\left\{ \begin{aligned} \sum_{k \in s} w_k (x_k - \hat{\mu}_{x,ee,CAL}) &= 0 \\ \sum_{k \in s} w_k ((x_k - \hat{\mu}_{x,ee,CAL})^2 - \hat{\sigma}_{x,ee,CAL}^2) &= 0. \end{aligned} \right.$$

Ce qui peut se réécrire :

Ce qui donne :

$$(\hat{\sigma}_{x,ee,CAL}^2 - \sigma_x^2)^{1/2} = \left(\frac{\hat{N}_\pi}{\hat{t}_{x,\pi}} - \frac{\hat{N}_{CAL}}{\hat{t}_{x,CAL}} \right) = O_p \left(\frac{\sqrt{n}}{1} \right).$$

et

$$\hat{\sigma}_{x,\pi}^2 - \sigma_x^2 = O_p \left(\frac{\sqrt{n}}{1} \right)$$

Ce résultat est intéressant car, à défaut d'un calage exact, on a un estimateur par calage de σ_x^2 qui est asymptotiquement plus précis que l'estimateur par substitution $\hat{\sigma}_{x,\pi}^2$. En effet, si on se place dans le cadre asymptotique habituellement utilisé en sondage et qu'on utilise la linéarisation des estimateurs complexes (Deville 1999), on a :

Ce qui correspond bien à l'expression de $\hat{\sigma}_{x,ee,CAL}^2$ donnée par l'équation (5).

$$\hat{N}_{CAL} \left(\frac{\hat{t}_{x,CAL}}{\hat{t}_{x,\pi}} - \frac{\hat{N}_\pi}{\hat{N}_{CAL}} \right)^2 - \hat{\sigma}_{x,ee,CAL}^2 = 0.$$

$$- \hat{N}_{CAL} (\sigma_x^2 - \hat{\sigma}_{x,ee,CAL}^2) = 0$$

$$\left(\frac{\hat{t}_{x,CAL}}{\hat{t}_{x,\pi}} - \frac{\hat{N}_\pi}{\hat{N}_{CAL}} \right) \left(2 \hat{t}_{x,CAL} \hat{N}_{CAL} - \hat{t}_{x,\pi} \hat{N}_{CAL} - \hat{t}_{x,CAL} \right)$$

$$- \hat{N}_{CAL} (\sigma_x^2 - \hat{\sigma}_{x,ee,CAL}^2) = 0$$

$$\left(\frac{\hat{t}_{x,CAL}}{\hat{t}_{x,\pi}} - \frac{\hat{N}_\pi}{\hat{N}_{CAL}} \right) \sum_{k \in s} w_k \left(2 x_k - \frac{\hat{t}_{x,\pi}}{\hat{t}_{x,CAL}} - \frac{\hat{N}_\pi}{\hat{N}_{CAL}} \right)$$

$$- \hat{N}_{CAL} (\sigma_x^2 - \hat{\sigma}_{x,ee,CAL}^2) = 0$$

$$\sum_{k \in s} w_k \left(\left(\frac{\hat{t}_{x,CAL}}{\hat{t}_{x,\pi}} - \frac{\hat{N}_\pi}{\hat{N}_{CAL}} \right) \left(2 x_k - \frac{\hat{t}_{x,\pi}}{\hat{t}_{x,CAL}} - \frac{\hat{N}_\pi}{\hat{N}_{CAL}} \right) \right)$$

En utilisant l'identité : $a^2 - b^2 = (a-b)(a+b)$, on a :

$$\sum_{k \in s} w_k \left(x_k - \frac{\hat{t}_{x,\pi}}{\hat{t}_{x,CAL}} \right)^2 - \left(x_k - \frac{\hat{t}_{x,\pi}}{\hat{t}_{x,CAL}} \right)^2 - \sigma_x^2 + \hat{\sigma}_{x,ee,CAL}^2 = 0.$$

on obtient :

- Si on fait la différence terme à terme de l'équation de calage approchée et de la deuxième équation estimante,

$$\left\{ \begin{aligned} \hat{\mu}_{x,ee,CAL} &= \frac{\sum_{k \in s} w_k x_k}{\sum_{k \in s} w_k} = \frac{\hat{t}_{x,CAL}}{\hat{t}_{x,\pi}} \\ \sum_{k \in s} w_k \left(x_k - \frac{\hat{t}_{x,CAL}}{\hat{t}_{x,\pi}} \right)^2 - \hat{\sigma}_{x,ee,CAL}^2 &= 0. \end{aligned} \right.$$

3.2 Calage dans le cas de paramètres définis par des équations estimentes

Soient $\mathbf{x}_i^k = (x_{(1)}^k, \dots, x_{(P)}^k)$, le vecteur des P variables auxiliaires connues sur s et η_x un paramètre complexe, également connu, défini par l'équation estimentante :

$$\sum_{k \in L} \Psi(\eta_x, \mathbf{x}_k) = 0.$$

Définition 2 : Dans le cas du calage sur le paramètre complexe η_x , les poids de calage sont obtenus par résolution du programme d'optimisation suivant :

$$\min_{\{w_k^k\}_{k \in s}} \sum_{k \in s} d(w_k, d_k)$$

sous contraintes :

$$\sum_{k \in s} w_k^k \Psi(\eta_x, \mathbf{x}_k) = 0.$$

Proposition 3 : Le calage sur un paramètre complexe η_x , défini par une equation estimentante, est équivalent à un calage sur le total de la nouvelle variable auxiliaire : $z_k = \Psi(\eta_x, \mathbf{x}_k)$, avec la contrainte de calage $\sum_{k \in s} w_k^k z_k = 0$.

Définition 3 : Un estimateur par calage du paramètre d'intérêt θ_y , noté $\hat{\theta}_{y,ee,CAL}$, est solution de l'équation estimentante sur s pondérée par les poids de calage $\{w_k^k\}_{k \in s}$:

$$\sum_{k \in s} w_k^k \Phi(\hat{\theta}_{y,ee,CAL}, \mathbf{y}_k) = 0.$$

Dans la plupart des cas la solution de l'équation estimentante est unique. La médiane est un exemple de paramètre qui peut offrir plusieurs solutions. On prend alors souvent l'infimum comme estimateur.

Proposition 4 : S'il y a unicité de la solution de l'équation : $\sum_{k \in s} w_k^k \Psi(\eta_{x,ee,CAL}, \mathbf{x}_k) = 0$, alors $\hat{\eta}_{x,ee,CAL} = \eta_x$.

Démonstration : η_x est solution de l'équation estimentante qui définit $\hat{\eta}_{x,ee,CAL}$. Puisqu'il y a une solution unique, on a $\hat{\eta}_{x,ee,CAL} = \eta_x$.

3.3 Calage sur une variance

L'objet de cette section est d'étudier le calage sur la variance σ_x^2 qui est un paramètre complexe plus compliqué que ceux vus précédemment. On montrera que lorsqu'on a uniquement ce paramètre comme information auxiliaire, on peut procéder à un calage approché qui fournit des poids de calage qui ont de meilleures propriétés que les poids d'échantillonnage.

Revenons donc sur le cas de la variance. La moyenne μ_x et la variance σ_x^2 sur U de la variable auxiliaire x peuvent être définies par deux équations estimentes sur U :

Tableau 1 Exemples de paramètres définis par des équations estimentes sur U

Paramètres	$\Phi(\theta_y, \mathbf{y}_k)$	Equations estimentes sur U
moyenne μ	$(y_k - \mu)$	$\sum_{k \in U} (y_k - \mu) = 0$
ratio $R = \mu_1 / \mu_2$	$(y_{(1)}^k - R y_{(2)}^k)$	$\sum_{k \in U} (y_{(1)}^k - R y_{(2)}^k) = 0$
médiane m	$(y_k \leq m - 1/2)$	$\sum_{k \in U} (1 - y_k \leq m - 1/2) = 0$

On peut donner également l'exemple du coefficient d'une régression logistique. Soient $y_{(1)}^k$ une variable dichotomique qui prend les valeurs 0 et 1 sur U , et $y_{(2)}^k$ une variable quantitative. On suppose que la valeur $y_{(1)}^k$ prises par $y_{(1)}^k$ pour l'unité k est une réalisation de la variable aléatoire $Y_{(1)}^k$ qui suit une loi de Bernoulli

$$\mathfrak{B}\left(1, p_k = \frac{1 + \exp(-\beta_0 y_{(2)}^k)}{1}\right).$$

Nous avons limité le nombre de paramètres à un, mais il serait aussi simple de considérer le cas multidimensionnel. Toutefois, il faudrait donner une définition des équations estimentes qui prennent en compte le cas des paramètres vectoriels.

Le paramètre qui nous intéresse est l'estimateur de β_0 , noté $\hat{\beta}$, calculé sur la population finie par la méthode du maximum de vraisemblance. L'équation estimentante de $\hat{\beta}$ sur U sera l'équation du maximum de vraisemblance. La log vraisemblance dans le cas de variables de Bernoulli vaut :

$$L(\beta) = \sum_{k \in L} y_{(1)}^k \ln(p_k) + \sum_{k \in L} (1 - y_{(1)}^k) \ln(1 - p_k).$$

On obtient facilement l'équation estimentante de $\hat{\beta}$ sur U :

$$\sum_{k \in U} y_{(2)}^k y_{(1)}^k \left(y_{(1)}^k - \frac{1 + \exp(-\beta y_{(2)}^k)}{1} \right) = 0.$$

L'équation estimentante sur s qui définit l'estimateur $\hat{\beta}_{ee,\pi}$ à partir des poids d'échantillonnage est :

$$\sum_{k \in s} d_k y_{(2)}^k y_{(1)}^k \left(y_{(1)}^k - \frac{1 + \exp(-\beta_{ee,\pi} y_{(2)}^k)}{1} \right) = 0.$$

L'équation estimentante n'est pas linéaire dans le paramètre ; $\beta_{ee,\pi}$ ne peut pas être exprimé comme une fonction simple des observations.

L'exemple de la régression logistique est très intéressant car il met en évidence qu'on n'a pas besoin de connaître l'expression de $\hat{\beta}_{ee,\pi}$ pour faire le calage. On verra dans la sous-section suivante qu'il suffit d'avoir le terme générique de l'équation estimentante sur

$$U, \Phi(\beta, \mathbf{y}_k) = y_{(2)}^k \left(y_{(1)}^k - \frac{1 + \exp(-\beta y_{(2)}^k)}{1} \right),$$

pour tout les $k \in s$.

(dans le cas où le plan de sondage permet d'estimer par-faitement la taille de la population N) :

$$\hat{\sigma}_{x,CAL}^2 = \frac{1}{N} \sum_{k \in s} w_k^2 \left(x_k - \left(\frac{\sum_{l \in s} w_l^2 x_l}{N} \right) \right)^2.$$

Trouver une nouvelle variable auxiliaire z n'est pas immédiat, dans la mesure où l'équation de calage initiale n'est pas linéaire par rapport au vecteur des poids. Nous reviendrons sur le cas de la variance à la section 3.3 de cet article.

Exemple du Ratio

Proposition 1 : Le calage sur un ratio est équivalent au calage sur le total de la nouvelle variable auxiliaire : $z_k = x_{k(1)}^k - R x_{k(2)}^k$.

L'équation de calage s'écrit :

$$t_{z,CAL} = t_z = 0.$$

Démonstration :

$$t_{z,CAL} = t_z$$

$$\Leftrightarrow \sum_{k \in s} w_k^2 (x_{k(1)}^k - R x_{k(2)}^k) = \sum_{k \in s} (x_{k(1)}^k - R x_{k(2)}^k)$$

$$\Leftrightarrow t_{x(1),CAL} - R x_{x(2),CAL} = t_{x(1),CAL} - R x_{x(2),CAL} = 0$$

$$\Leftrightarrow \frac{t_{x(1),CAL}}{t_{x(2),CAL}} = R$$

$$\text{i.e., } R_{x,CAL} = R_x.$$

Fonction d'un ratio de combinaisons linéaires de totaux

Soit η_x un paramètre complexe qui est une fonction bijective d'un ratio de combinaisons linéaires de totaux :

$$(1) \quad \eta_x = h \left(\frac{\alpha' \cdot t_x}{\beta' \cdot t_x} \right)$$

avec $\alpha' = (\alpha_1, \dots, \alpha_p)$ et $\beta' = (\beta_1, \dots, \beta_p)$ des vecteurs de coefficients réels de taille p , et $t_x^i = (t_{x(1)}^i, \dots, t_{x(p)}^i)$.

Proposition 2 : Réaliser un calage sur le paramètre complexe η_x défini par la fonction (1) est équivalent à caler sur le total de la nouvelle variable auxiliaire :

$$z_k = (\alpha' - h^{-1}(\eta_x) \beta') \cdot x_k$$

$$t_{z,CAL} = \sum_{k \in s} w_k^2 z_k = t_z = 0.$$

avec l'équation de calage :

Nous allons voir dans la suite de cet article que la méthode des équations estimantes offre une autre approche pour exhiber la (les) nouvelle(s) variable(s) auxiliaire(s) z .

3. Paramètre défini par une équation estimante

3.1 Principe d'estimation par équation estimante

Certains paramètres θ_y se définissent, ou peuvent se définir, comme solution d'une fonction implicite appelée *équation estimante* sur U (Godambe et Thompson 1986), i.e. :

$$\sum_{k \in U} \Phi(\theta_y, y_k) = 0$$

avec $y_k^i = (y_{(1)}^k, \dots, y_{(p)}^k)$ le vecteur des valeurs prises par les variables d'intérêt pour l'individu k . Dans ce contexte, on définit un estimateur de θ_y à partir de l'échantillon s , noté $\hat{\theta}_{y,est}$, qui est la solution de l'équation estimante sur s (voir, entre autres, Hidiroglou, Rao et Yung 2002) :

$$\sum_{k \in s} d_k \Phi(\hat{\theta}_{y,est}, y_k) = 0.$$

Il s'agit simplement de la fonction $f(\cdot, \dots, \cdot)$ dans laquelle les totaux $t_{y^{(g)}}^{x^{(1)}, \dots, x^{(p)}}$ ont été remplacés par leur estimateur Horvitz-Thompson $\hat{t}_{y^{(g)}, \pi}^{x^{(1)}, \dots, x^{(p)}} = \sum_{k \in s} d_k^k y_{(g)}^{(k)}(s)$ (Särndal, Swensson and Wretman 1992). On peut qualifier cet estimateur d'estimateur par substitution.

Soient $x^{(1)}, \dots, x^{(p)}, P$ variables auxiliaires, connues sur s et $t_{y^{(g)}}^{x^{(1)}, \dots, x^{(p)}}$, les totaux sur U de ces variables auxiliaires qui sont également connus. Pour un individu k , on note $\mathbf{x}_k' = (x_{(1)}^k, \dots, x_{(p)}^k)$ le vecteur des valeurs prises par les variables auxiliaires sur k .

L'estimateur par calage de θ_y est :

$$\hat{\theta}_{y, \text{CAL}} = f(\hat{t}_{y^{(1)}, \text{CAL}}, \dots, \hat{t}_{y^{(p)}, \text{CAL}})$$

avec $\hat{t}_{y^{(g)}, \text{CAL}} = \sum_{k \in s} w_k^k y_{(g)}^{(k)}$, et une série de poids $\{w_k^k\}_{k \in s}$, dits poids de calage (que l'on devrait noter $w_k^k(s)$, puisque ces poids dépendent de l'échantillonnage), obtenus par résolution du programme d'optimisation suivant :

$$\min \sum_{k \in s} d(w_k^k, d_k^k)$$

sous contraintes :

$$\begin{cases} \hat{t}_{y^{(1)}, \text{CAL}} = t_{y^{(1)}} \\ \dots \\ \hat{t}_{y^{(p)}, \text{CAL}} = t_{y^{(p)}} \end{cases}$$

$d(\cdot, \cdot)$ est une pseudo-distance, i.e., une fonction qui mesure l'écart du poids de calage au poids d'échantillonnage (à la différence d'une distance, la pseudo-distance n'est pas nécessairement symétrique sur ses deux arguments). Le programme se résout à l'aide d'un Lagrangien. Dans le cas où la distance utilisée est la distance du χ^2 (i.e., $d(w_k^k, d_k^k) = (1/2)(w_k^k - d_k^k)^2/d_k^k$, on trouve comme solution : $w_k^k = d_k^k(1 + \mathbf{x}_k' \lambda)$ (où λ est le vecteur de taille P des multiplicateurs de Lagrange).

2.2 Calage sur un paramètre complexe η_x

Définition 1 : Soient $x^{(1)}, \dots, x^{(p)}, P$ variables auxiliaires, connues sur s et $\eta_x = g(t_{x^{(1)}}^{x^{(1)}, \dots, x^{(p)}}, \dots, t_{x^{(p)}}^{x^{(1)}, \dots, x^{(p)}})$ un paramètre complexe, fonction des totaux de ces variables auxiliaires, également connu.

Dans le cas du calage sur le paramètre complexe η_x , les poids de calage sont obtenus par résolution du programme d'optimisation suivant :

$$\min \sum_{k \in s} d(w_k^k, d_k^k)$$

sous contraintes :

$$\hat{\eta}_{x, \text{CAL}} = g(\hat{t}_{x^{(1)}, \text{CAL}}, \dots, \hat{t}_{x^{(p)}, \text{CAL}}) = \eta_x.$$

Les totaux $t_{x^{(g)}}^{x^{(1)}, \dots, x^{(p)}}$ n'ont pas besoin d'être connus, seul le paramètre complexe η_x doit l'être.

Prenons l'exemple du ratio

$$R_x = \frac{t_{x^{(1)}}^{x^{(1)}, \dots, x^{(p)}}}{\sum_{k \in U} x_{(1)}^k} = \frac{t_{x^{(1)}, \pi}^{x^{(1)}, \dots, x^{(p)}}}{\sum_{k \in U} x_{(1)}^k}.$$

L'estimateur par calage de R_x est de la forme :

$$\hat{R}_{x, \text{CAL}} = \frac{\sum_{k \in s} w_k^k x_{(1)}^k}{\sum_{k \in s} w_k^k x_{(2)}^k}.$$

L'équation de calage dans le cas d'un calage sur le ratio s'écrit :

$$\hat{R}_{x, \text{CAL}} = \frac{\sum_{k \in s} w_k^k x_{(1)}^k}{\sum_{k \in s} w_k^k x_{(2)}^k} = R_x$$

R_x est une information auxiliaire connue, comme l'est habituellement le total des variables auxiliaires. Ce cas de figure peut se produire lorsqu'on a des proportions bien connues et qui par exemple sont stables dans le temps, mais qu'on ne connaît pas bien les totaux du dénominateur et du numérateur.

On a présenté le cas du calage sur un seul paramètre complexe, mais on voit qu'il est facile de caler sur plusieurs paramètres complexes. Il y a alors autant de contraintes que de paramètres de calage.

2.3 Cas simples où le calage sur paramètre complexe peut se ramener au calage sur un total

Il n'est pas évident de savoir d'emblée si l'équation de calage sur un paramètre complexe va pouvoir s'écrire sous la forme d'une équation de calage sur un total. En d'autres termes, il n'est pas toujours trivial de trouver une « nouvelle » variable auxiliaire z , associée au paramètre complexe, sur le total de laquelle on va caler. Par exemple, cela est assez immédiat pour tous les moments d'une variable auxiliaire x (on suppose ici que le plan de sondage permet d'estimer parfaitement la taille de la population N). Si $\mu_{x^m}^{x^m} = N^{-1} \sum_{k \in U} x_k^m$ est une information auxiliaire, alors il suffit de prendre $z_k = x_k^m / N$ et de caler sur $\mu_{x^m}^{x^m}$: $\sum_{k \in s} w_k^k x_k^m / N = \mu_{x^m}^{x^m}$. Si on veut caler sur la variance et la moyenne de la variable x en ayant μ_x et σ_x^2 comme information auxiliaire, alors on peut utiliser les deux nouvelles variables auxiliaires :

$$z_{(1)}^k = \frac{x_k}{N}$$

et

$$z_{(2)}^k = \frac{N}{(x_k - \mu_x)^2}.$$

En revanche, si on ne connaît pas μ_x , mais qu'on a σ_x^2 dans l'information auxiliaire et qu'on veut caler sur cette variance, alors les choses se compliquent. Pour s'en convaincre, il faut écrire l'estimateur par substitution de σ_x^2

Nous concluons cette section par un exemple succinct démontrant que la PPT n'est pas *toujours* meilleure que l'estimateur par quotient. Bien que les données dans le tableau 2 pour une population de cinq unités soient artificielles, un modèle de données de ce genre pourrait survenir dans les succursales financières où de très petites entreprises financières peuvent grandir très vite à l'égard de certaines variables financières. Cette forte variabilité des taux de croissance des petites entreprises entraîne une faible valeur pour δ . Pour un échantillon EAS où $n = 2$ pour les cinq unités au tableau 2, la variance de l'estimateur par quotient est de 211 selon (12); les simulations donnent $EQM_{(sim)}^R = 323$. C'est beaucoup moins que la variance de 557 déterminée dans une simulation comportant 80 000 échantillons PPT aléatoires de taille $n = 2$. La formule (3) agencée à (9) donne le même résultat : 557. Ce serait également la bonne variance si l'échantillon avait été tiré selon Brewer (1963a) ou Durbin (1967). La formule (11), basée sur (10), donne une valeur légèrement différente : 556.

La régression (19) avec les données du tableau 2 donne $\beta = -3,0$, et donc $\delta = -1,0$. Conformément aux résultats de la sous-section 3.3, cette faible valeur $\delta = -1,0$ explique pourquoi \hat{Y}_{PPT}^R est moins efficace que \hat{Y}_R^R dans cet exemple. De plus, l'estimateur direct ordinaire $\hat{N}\hat{Y}_s^R$ d'un échantillon EAS a une variance de 356, qui est encore plus petite ici que la variance dans l'échantillonnage PPT aléatoire, sachant que \hat{Y}_s^R est la moyenne de l'échantillon de X_i . Par conséquent, pour ce type de données, l'estimateur par quotient est la meilleure solution. Rappelons que l'estimateur par quotient a une plus petite variance que $\hat{N}\hat{Y}_s^R$ lorsque $b > Y/2X$, où b est la pente de la droite de régression de Y_i sur X_i et une constante ($i = 1, \dots, N$); voir Klottnernus (2003, page 117). Ainsi, les données X_i ($= X_i/Z_i$) dans le tableau 2 ne présentent certainement pas une tendance stable.

Le présent document compare la variance de l'estimateur HT \hat{Y}_{PPT}^R à partir d'un échantillon PPT aléatoire avec la variance de l'estimateur classique par quotient \hat{Y}_R^R à partir

5. Sommaire

<i>i</i>	Taux de croissance	Taille
1	200 %	0,0455
2	33 %	0,1364
3	75 %	0,1818
4	33 %	0,2727
5	62 %	0,3636

Tableau 2
Taux de croissance des biens (Z_i) et taille (X_i) de cinq établissements

Contre-exemple

Annexe A

Les opinions exprimées dans l'article sont celles de l'auteur et ne correspondent pas nécessairement à la politique du bureau central de la statistique des Pays-Bas. L'auteur tient à remercier Peter-Paul de Wolf, Sander Scholhuis, le rédacteur en chef adjoint et deux réviseurs anonymes pour leurs suggestions et leurs corrections utiles.

Remerciements

Jours négligeable.

Lorsque les tendances des données des variables x et z ($= y/x$) sont telles que $p_z < -1/(N-1)$, on peut démontrer dans certaines conditions peu contraignantes que \hat{Y}_{PPT}^R est plus efficace que \hat{Y}_R^R lorsque n et N sont assez grands, à condition que X_i et Z_i soient sans corrélation. En vertu du modèle (15) où $E(e_i^2) = \sigma^2 X_i^8$, on constate que $p_z < -1/(N-1)$ lorsque $\delta > 1$. Par conséquent, pour ce type de données, il faut préconiser \hat{Y}_{PPT}^R . De plus, il ressort de (14) et (16) que pour $\delta = 2$, l'efficacité relative de l'échantillonnage PPT comparativement à celle de l'estimateur par quotient augmente parallèlement au CV^x . En outre, il faut privilégier \hat{Y}_R^R lorsque les données correspondent à un modèle où $\delta < 1$. Ces résultats sont confirmés de façon empirique avec une étude de simulation axée sur deux différents ensembles de données. Lorsque le modèle (15) ne s'applique pas, l'efficacité relative de \hat{Y}_{PPT}^R est donnée par (14), à condition que n soit grand et que N soit relativement plus grand. En pratique, le p_z inconnu dans (14) est remplacé par \hat{p}_{z9} . Le fait que $n \ll N$ ne signifie pas nécessairement que le facteur $(n-1)p_z$ dans (3) est toujours négligeable.

Les équations (5) et (7) ne peuvent pas toujours être utilisées pour l'échantillonnage PPT aléatoire lorsque n et N sont du même ordre, tandis que X_i et Z_i sont corrélés. Pour le constater, supposons une population U composée de deux groupes, U_1 et U_2 , avec des moyennes de \bar{X}_1 et \bar{X}_2 , respectivement. Les deux tailles de strates sont $N/2$. Supposons que s est un échantillon PPT aléatoire de taille $n = 3N/4$ tiré de l'ensemble de la population U . Supposons que X_i est tel que

$$\pi_i = nX_i' = \begin{cases} 1 & \text{si } i \in U_1 \\ 0,5 & \text{si } i \in U_2 \end{cases}$$

De toute évidence, le groupe 1 ne contribue pas à la variance. Les unités sélectionnées en s de U_2 constituent un

moyenne (EQM) et le biais de \hat{Y}_R , pour obtenir $EQM_R^{(sim)} = 108$ et un biais relativement petit de 0,7. Voilà qui confirme la conjecture que (12) donne une sous-estimation importante de la variance réelle ; voir Cochran (1977). Par conséquent, pour des échantillons modérés, la véritable valeur de $Eff_{P/R}$ pourrait être quelque peu plus élevée que (14) le suggère.

De plus, il convient de souligner que la formule simplifiée (10) pour p_z combinée à (3) donne presque le même résultat $var(\hat{Y}_{PPT}^2) = 30,7$, même si $N = 70$, n'est pas très grand. La variance PPT avec remplacement aurait été de 43,8. Par conséquent, la réduction de la variance pour l'échantillonnage PPT aléatoire est supérieure à 30 %, même si la fraction d'échantillonnage n/N est bien plus petite. Selon (18), la formule (12) où $n_{EAS} = 26$ donne à peu près le même résultat que (3) où $n_{PPT} = 9$; nota : $p_z = -0,042$. Par conséquent, la taille des échantillons diverge

Tableau 1
Variations des prix (Z_i) et tailles (X_i) de 70 établissements

i	variation des prix	taille	i	variation des prix	taille
1	-18,4 %	0,0608	36	34,8 %	0,0427
2	-16,0 %	0,0784	37	13,1 %	0,0121
3	3,3 %	0,0762	38	31,7 %	0,0351
4	12,5 %	0,0100	39	-24,8 %	0,0074
5	0,0 %	0,0029	40	55,3 %	0,0009
6	8,3 %	0,0006	41	40,5 %	0,0066
7	-39,0 %	0,0182	42	34,6 %	0,0022
8	-25,1 %	0,0020	43	1,7 %	0,0001
9	1,1 %	0,0040	44	0,0 %	0,0039
10	4,4 %	0,0066	45	3,9 %	0,0304
11	-4,9 %	0,0039	46	25,4 %	0,0209
12	-8,9 %	0,0070	47	25,6 %	0,0062
13	-7,0 %	0,0148	48	0,0 %	0,0033
14	-15,0 %	0,0108	49	-0,3 %	0,0019
15	-10,7 %	0,0087	50	66,6 %	0,0346
16	-9,0 %	0,1079	51	0,0 %	0,0039
17	-11,3 %	0,0247	52	-2,9 %	0,0007
18	10,6 %	0,0024	53	15,8 %	0,0011
19	-23,2 %	0,0001	54	0,0 %	0,0026
20	-25,4 %	0,0001	55	0,0 %	0,0018
21	-80,7 %	0,0002	56	11,6 %	0,0057
22	13,4 %	0,0005	57	0,0 %	0,0042
23	-42,5 %	0,0010	58	0,0 %	0,0236
24	-34,8 %	0,0014	59	-1,5 %	0,0015
25	-30,0 %	0,0126	60	0,0 %	0,0003
26	8,0 %	0,0530	61	11,7 %	0,0067
27	0,0 %	0,0208	62	0,0 %	0,0012
28	2,1 %	0,0119	63	0,8 %	0,0040
29	11,3 %	0,0208	64	2,0 %	0,0009
30	0,7 %	0,0322	65	2,3 %	0,0018
31	9,5 %	0,0447	66	4,7 %	0,0026
32	11,5 %	0,0018	67	0,9 %	0,0064
33	5,8 %	0,0174	68	-1,0 %	0,0309
34	-6,9 %	0,0197	69	-0,5 %	0,0005
35	0,0 %	0,0124	70	0,0 %	0,0006

générale, en supposant que $(n-1)/n \approx 1$, on constate à partir de (3), (10) et (12) qu'un estimateur par quotient d'un EAS de taille n^{EAS} est aussi efficace qu'un échantillon PPT de taille n^{PPT} lorsque

$$n^{\text{EAS}} = -n^{\text{PPT}} \rho^z N. \quad (18)$$

3.3 Efficacité de X^{PPT} pour $\delta < 1$ par rapport à $\delta \geq 1$

Un autre cas spécial est $\delta = 1$. À partir de (16), $\rho^z = -1/N$ lorsque $\delta = 1$. Ensuite, on constate d'après (14) qu'en vertu du modèle (15), $\text{Eff}_{P/R}^{\text{eff}} = 1 + O(N^{-1})$, à condition que $n/N < f_0 < 1$, lorsque $N \rightarrow \infty$ sans égard à la valeur de CV^x . De plus, on peut démontrer que $\text{Eff}_{P/R}^{\text{eff}}$ est une fonction croissante de δ . Cette théorie est prouvée ci-après au Lemme 1. Par conséquent, pour $\delta < 1$, l'estimateur PPT aléatoire est moins efficace que l'estimateur par quotient, alors que pour $\delta > 1$, l'estimateur PPT aléatoire est plus efficace que l'estimateur par quotient.

Lemme 1. Supposons que $\text{Eff}_{P/R}^{\text{eff}}$ et que ρ^z sont définis par (14) et (16), respectivement. Si $V^2 > 0$, alors $\text{Eff}_{P/R}^{\text{eff}}$ est une fonction croissante monotone de δ .

Preuve. Écrivez ρ^z à partir de (16) comme une moyenne pondérée de X_i (négatif).

$$\rho^z = -n(\delta) = -\sum_{i \in U} w_i X_i,$$

où

$$w_i = \frac{X_i^{\delta-1}}{\sum_{i \in U} X_i^{\delta-1}} \quad [\text{Notons que } \mu_x = n(2)].$$

Supposons que $X_i > X_j$ ($i \neq j$), et définissons $h(\delta)$ comme $w_i/w_j = (X_i/X_j)^{\delta-1}$. Étant donné que $h(\delta)$ augmente en fonction de δ , le poids du grand X_i augmente comparativement à celui de X_j lorsque δ augmente. Par conséquent, $n(\delta)$ augmente et ρ^z décroît en fonction de δ . Il suffit donc de démontrer que $\text{Eff}_{P/R}^{\text{eff}}$ diminue en fonction de ρ^z . En écrivant (14) comme

$$\text{Eff}_{P/R}^{\text{eff}} = \frac{-(N-n)}{p^z_{-1} + (n-1)},$$

on constate que $\text{Eff}_{P/R}^{\text{eff}}$ diminue effectivement en fonction de ρ^z . Voilà qui conclut la preuve.

3.4 Une structure de rechange parmi les perturbations

Enfin, supposons que la variance des perturbations à (15) est de la forme suivante :

$$\text{var}(e_i) = c_1 X_i + c_2 X_i^2 \quad (0 < c_1, c_2 \leq 1).$$

Voir Kott (1988). Pour ce cas, nous obtenons par analogie avec (16)

$$\rho^z = -\sum_{i \in U} w_i X_i,$$

où

$$w_i = \frac{\sum_{i \in U} (1 + \phi X_i)}{1 + \phi X_i}, \quad \text{et } \phi = c_2/c_1$$

lorsque $\phi = 0$, $\rho^z = -1/N$. Ainsi, lorsque $c_2 = 0$, l'efficacité de l'échantillonnage PPT est seulement équivalente à l'estimateur par quotient ordinaire de l'EAS. Dans la même veine que la preuve du Lemme 1, on peut démontrer que ρ^z diminue en fonction de ϕ , tandis que $\text{Eff}_{P/R}^{\text{eff}}$ augmente en fonction de ϕ . Par conséquent, dans ce cas-ci, l'estimateur PPT aléatoire est toujours plus efficace que l'estimateur par quotient lorsque c_2 est positif.

4. Une application de l'Indice des prix à la production

L'Indice des prix à la production (IPP) des Pays-Bas est basé sur environ 2 500 indices de prix de marchandises, organisés par type de produit. L'indice de prix pour un produit précis peut être écrit comme suit

$$Y = \sum_{i \in U} X_i Z_i,$$

où Z_i est la variation des prix pour ce produit de l'établissement i par rapport à la période de base, tandis que X_i représente les ventes relatives de ce produit par l'établissement i pendant la période de base (rappelons que $\sum X_i = 1$).

Dans l'exemple donné ici, nous examinons les variations de prix de 70 établissements pour le produit *Métal de base* en décembre 2005 par rapport à décembre 2004 ; voir le tableau 1. Nous comparons la variance de l'estimateur par quotient d'un échantillon EAS à la variance de l'estimateur HT d'un échantillon PPT aléatoire lorsque $n = 9$. En appliquant (12) à ces données, on obtient $\text{var}(Y^R) = 101$. Si l'échantillon avait été tiré avec remplacement, la variance aurait été de 116. En appliquant (3) et (9) pour un échantillon PPT aléatoire, on obtient $\text{var}(Y^{\text{PPT}}) = 29,9$. Ce résultat tient compte de γ et est proche du résultat $V^{\text{PPT}} = 29,2$ d'une expérience de simulation comprenant 80 000 échantillons PPT aléatoires de taille $n = 9$ à partir de l'ensemble de 70 établissements. Ainsi, $\text{Eff}_{P/R}^{\text{eff}} = 3,5$. Étant donné que la formule (12) pour $\text{var}(Y^R)$ est seulement asymptotiquement sans biais, nous avons également effectué des simulations pour évaluer l'erreur quadratique

Maintenant, à partir de (3), on constate que $s_z^2/(1-p_z)$ est un estimateur sans biais pour σ_z^2 . Lorsque p_z est très petit, le terme $(1-p_z)$ peut être omis. Lorsque n est assez grand, le ratio p_z de (9) peut être estimé comme suit

$$\hat{p}_{z9} = -\frac{\sum_{i \in S} X_i (Z_i - \bar{z})^2 / \hat{\gamma} (1 - 2X_i)}{\sum_{i \in S} (Z_i - \bar{z})^2},$$

où

$$\hat{\gamma} = \frac{1}{1} + \frac{2n}{1} \sum_{i \in S} \frac{1 - 2X_i}{1}.$$

Comme $\hat{\gamma} \geq 1$ et $X_i \leq 1/n$, nous obtenons $\hat{p}_{z9} \geq -1/(n-2)$. Pour le biais d'un ratio estimatif lorsque n est petit, voir Cochran (1977, page 160).

De même, l'élément p_z dans (10) peut être estimé comme suit

$$\hat{p}_{z10} = -\frac{\sum_{i \in S} X_i (Z_i - \bar{z})^2}{\sum_{i \in S} (Z_i - \bar{z})^2} \geq \frac{n}{-1} > \frac{n}{n-1}.$$

Donc, en remplaçant σ_z^2 et p_z dans (3) par $s_z^2/(1-\hat{p}_{z10})$ et \hat{p}_{z10} , respectivement, on obtient un estimateur de variance non négatif avec une probabilité de 1. On peut en dire autant pour \hat{p}_{z9} , lorsque $X_i \leq 1/(n+1)$. L'estimateur pour $\text{var}(Y^{\text{ppt}})$ ainsi obtenu devient

$$\widehat{\text{var}}_p(Y^{\text{ppt}}) = \frac{\{1 + (n-1)\hat{p}_{z9}\}s_z^2}{n(1-\hat{p}_{z9})}.$$

En outre, pour les valeurs modérées de N , l'estimateur \hat{p}_{z9} a probablement de meilleures propriétés que \hat{p}_{z10} parce que les éléments π_{ijk}^{JK} sous-jacents à (9) satisfont exactement aux restrictions de deuxième ordre sans égard aux valeurs de n et de N .

3. Efficacité de Y^{ppt} pour grands n et N

3.1 Formules d'efficacité

Comme $X = 1$, l'estimateur par quotient pour Y devient

$$\hat{Y}_R = \frac{\bar{Y}_S}{\bar{X}_S} = \frac{\sum_{i \in S} X_i Z_i}{\sum_{i \in S} X_i}.$$

Lorsque n est assez grand, l'approximation souvent utilisée pour sa variance est la suivante

$$\text{var}(\hat{Y}_R) = \frac{n(N-n)}{N(N-1)} \sum_{i \in U} X_i^2 (Z_i - Y)^2. \quad (12)$$

À partir de (3) et (12), on peut constater que l'efficacité de Y^{ppt} par rapport à Y_R peut être écrite comme suit

$$\text{Eff}_{p/R} = \frac{\text{var}(Y^{\text{ppt}})}{\text{var}(\hat{Y}_R)} = \frac{(N-n) \sum_{i \in U} X_i^2 (Z_i - Y)^2}{\{1 + (n-1)p_z\} \sigma_z^2}, \quad (13)$$

Souignons qu'en remplaçant $n = n^{\text{ppt}}(1 + CV_x^2)$ dans (12), on obtient à peu près le même résultat qu'en (3) et (10) en remplaçant n par n^{ppt} . Ainsi, lorsque $CV_x^2 = 1.5$, l'échantillonnage PPT aléatoire dont la taille de l'échantillon $n^{\text{ppt}} = 100$ est aussi efficace que l'estimateur par quotient d'un EAS de taille $n^{\text{EAS}} = 325$. D'une manière plus

grand et que $n/N < f_0 < 1$.
cette hypothèse se maintient lorsque N est suffisamment supposant que $(N-n+1) \approx (N-n)$; souignons que à l'échantillonnage EAS et manifestement, $\text{Eff}_{p/R} = 1$, en Lorsque $CV_x^2 = 0$, l'échantillonnage PPT aléatoire équivalait aléatoire est élevée lorsque la variabilité de X_i est élevée. Par conséquent, si $\delta = 2$, l'efficacité de l'échantillon PPT

$$\text{Eff}_{p/R} = \frac{(N-n)(1+CV_x^2)}{(N-n)(1+CV_x^2)}.$$

de X_i . Lorsque l'on remplace (17) dans (14), on obtient où $\bar{X} = 1/N$ et $CV_x^2 = V_x/\bar{X}$ est le coefficient de variation

$$\frac{1}{N} \sum_{i \in U} X_i^2 = \bar{X}^2 + \bar{X}^2 = \bar{X}^2(1 + CV_x^2),$$

parce que

$$p_z = -\frac{1}{N}(1 + CV_x^2), \quad (17)$$

Comme $\delta = 2$, (16) donne $p_z = -\sum_{i \in U} X_i^2 = -\mu_x$, ce qui peut également être écrit comme suit

3.2 Efficacité de Y^{ppt} si $\delta = 2$

Dans les prochaines sous-sections, nous examinons de plus près la relation entre δ et l'efficacité de Y^{ppt} .

$$p_z = -\frac{\sum_{i \in U} X_i^2}{\sum_{i \in U} X_i^{\delta-1}}. \quad (16)$$

modèle, ce qui donne

et le dénominateur dans (10) par les valeurs attendues du grand, nous pouvons remplacer X_i , ainsi que le numérateur lorsque $N \rightarrow \infty$. De plus, lorsque N est suffisamment (14) se maintiennent lorsque n et N sont du même ordre qu'en supposant un modèle comme (15), (7) et donc (10) et Brewer (1963b), Brewer et Donadio (2003) ont démontré Koit (1988), δ se trouve souvent entre 1 et 2. Voir aussi $E(u_i) = 0, E(u_i^2) = \sigma^2 X_i^{\delta-2}$, et $E(u_i u_j) = 0 (i \neq j)$. D'après conséquent, pour Z_i , nous obtenons $Z_i = \mu + u_i$, où où $E(e_i) = 0, E(e_i^2) = \sigma^2 X_i^{\delta}$, et $E(e_i e_j) = 0 (i \neq j)$. Par

$$Y_i = \mu X_i + e_i, \quad (15)$$

au modèle :

Supposons maintenant que les observations Y_i satisfont

$$\text{Eff}_{p/R} = \frac{1 + (n-1)p_z}{-(N-n)p_z}. \quad (14)$$

(10) et (13), on obtient

en supposant que $N/(N-1) \approx 1$. Lorsque l'on combine

Étant donné que (5) et (7) entraînent des approximations pour p_z dans l'échantillonnage PPT aléatoire qui sont $p_z\{1+o(1)\}$ lorsque $N \rightarrow \infty$, avec $n = o(N)$, (5) peut être utilisé pour calculer p_z dans la pratique lorsque $n \ll N$ et que N est grand. Pour faciliter l'exposé, on présume ici qu'il y a une constante positive c , de sorte que $p_z < -c/N$.

Voir aussi Kott (2005, page 436), qui explique l'estimation de la variance sous l'échantillonnage PPT lorsque $n = O(N^{2/3})$.

Supposons que $\gamma = 1 + \mu_x + O(1/N^2)$ et que $\mu_x = O(1/N)$ (qui suit selon les conditions du théorème 1 ci-après). Il n'est pas difficile de voir que, après élimination des termes $O(1/N)$, c_i dans (6) est identique à c_{iHR} = $(n-1) / \{n(1 + \mu_x - 2X_i)\}$. Cette dernière expression est l'équation (11) de Brewer et Donadio, qui est basée sur π_{iHR} en (7).

L'approche proposée ici est un peu différente de Knothnerus (2003). D'abord, réécrivez (5) comme suit

$$\pi_{ijk} = n(n-1) \frac{\gamma}{1/2} \frac{X_i X_j}{1-2X_i} \left(\frac{1}{1/2} + \frac{1-2X_j}{1/2} \right). \quad (8)$$

En remplaçant (8) dans (4), nous obtenons une nouvelle approximation simple pour p_z :

$$p_z = \sum_{i \in U} \sum_{j \in U, j \neq i} \frac{\gamma}{1/2} \frac{X_i X_j}{1-2X_i} \left(\frac{1}{1/2} + \frac{1-2X_j}{1/2} \right) \left(\frac{\sigma_z}{Z_i - Y} \right) \left(\frac{\sigma_z}{Z_j - Y} \right)$$

$$= \sum_{i \in U} \sum_{j \in U, j \neq i} \frac{\gamma}{1-2X_i} \left(\frac{1}{1-2X_i} + \frac{1-2X_j}{1-2X_i} \right) \left(\frac{\sigma_z}{Z_i - Y} \right) \left(\frac{\sigma_z}{Z_j - Y} \right)$$

$$= 0 - \sum_{i \in U} \frac{\gamma}{1-2X_i} \left(\frac{\sigma_z}{Z_i - Y} \right). \quad (9)$$

À la deuxième ligne, nous avons utilisé l'égalité $\sum_{i,j} m_{ij} \gamma_i = \sum_{i,j} m_{ij} \gamma_j$ lorsque $m_{ij} = m_{ji}$. À la dernière ligne, nous avons utilisé $\sum_{j \in U} X_j (Z_j - Y) = 0$.

Ensuite, supposons que \bar{X} indique la moyenne de la population de X_1, \dots, X_N et définissons σ_x^2 et V_x^2 par

$$\sigma_x^2 = \sum_{i \in U} X_i (X_i - \mu_x)^2,$$

$$V_x^2 = \sum_{i \in U} (X_i - \bar{X})^2 / N,$$

respectivement. Dans le théorème suivant, (9) est encore plus simplifiée.

Théorème 1. On suppose que $(Z_i - Y)/\sigma_z = O(1)$, lorsque $N \rightarrow \infty$ et qu'il y a des constantes positives c et C , de sorte que $V_x/\bar{X} < c$, $\sigma_x/\mu_x < c$ et $0 < X_i < C < 1/2$. Ainsi, pour de grandes valeurs de N et $n \ll N$,

$$p_z = - \frac{\sum_{i \in U} X_i^2 (Z_i - Y)^2}{\sum_{i \in U} X_i (Z_i - Y)^2} \left\{ 1 + O\left(\frac{1}{N}\right) \right\} + O\left(\frac{1}{N^2}\right). \quad (10)$$

Preuve. Comme $\bar{X} = 1/N$, on conclut des suppositions qui précèdent que la moyenne pondérée $\mu_x = \sum X_i^2 = N(V_x^2 + \bar{X}^2)$ est de l'ordre de $1/N$ et donc, $\sigma_x = O(1/N)$. Comme $(1-2X_i)^{-1} = 1 + 2X_i + O(X_i^2)$ pour $0 < X_i < C < 1/2$, p_z de (9) peut être écrit lorsque $N \rightarrow \infty$, comme

$$p_z = - \sum_{i \in U} \frac{\gamma}{1} \frac{X_i^2}{Z_i - Y} \left(\frac{\sigma_z}{Z_i - Y} \right) + \frac{\gamma}{1} O\left(\sum_{i \in U} X_i^3\right),$$

où $\sum_{i \in U} X_i^3 = \sigma_x^3 + \mu_x^3 = O(N^{-2})$, et

$$\gamma = \frac{1}{1} + \frac{2}{1} \sum_{i \in U} X_i \{1 + 2X_i + O(X_i^2)\}$$

$$= 1 + \mu_x + O\left(\frac{1}{N^2}\right) = 1 + O\left(\frac{1}{N}\right).$$

dont suit (10). Voilà qui conclut la preuve.

En substituant (10) dans (3), nous obtenons

$$\text{var}(Y^{\text{ppt}}) = \frac{\sigma_z^2}{n} - \frac{n}{n-1} \sum_{i \in U} X_i^2 (Z_i - Y)^2$$

$$= \frac{1}{n} \sum_{i \in U} X_i^2 \{1 - (n-1)X_i\} (Z_i - Y)^2, \quad (11)$$

une équation qui est également proposée par Hartley et Rao (1962). Il convient de souligner que l'approximation (10) découle également directement de la substitution de l'approximation simple $\pi_{iAP} = n(n-1)X_i X_j$ dans (4). De même, l'utilisation de π_{iHR} donne lieu à une expression semblable à (9), et donc, à (10). De plus, l'utilisation directe de π_{iAP} à (1) ou (2) pour le cas d'EAS, où $X_i = X_j = 1/N$ pourait entraîner des erreurs de plus de 100 % pour les populations où $\bar{X} = V_x^2$; voir Knothnerus (2003, pages 274-276). Donc, (1) et (2) sont plus sensibles aux petites erreurs dans π_{ij} que (3) et (4). Qui plus est, soulignons que lorsque n est petit à un point tel que $|np_z| \ll 1$, nous pouvons établir que $p_z = 0$, ce qui donne la formule de variance avec remplacement de Hansen et Hurwitz (1943).

Pour estimer (3) au moyen de p_z , comme précédemment, dénotons par z_{s1} une observation choisie au hasard de s . Nous obtenons alors

$$\sigma_z^2 = \text{var}(z_{s1}) = \text{var}\{E(z_{s1}|s)\} + E\{\text{var}(z_{s1}|s)\} \\ = \text{var}(\bar{z}_s) + E\left(\frac{n}{n-1} s_z^2\right),$$

où

$$s_z^2 = \frac{1}{n-1} \sum_{i=1}^n (Z_i - \bar{z}_s)^2.$$

Bien que les expressions exactes pour π_{ij} selon l'échantillonnage PPT aléatoire soient disponibles, elles peuvent être lourdes lorsque N est grand. Pour une expression exacte, voir Connor (1966) et pour une modification, voir Hidiroglou et Gray (1980). Nous utilisons ici une approximation proposée par Knottnerus (2003, page 197) :

$$\pi_{ijk} = n(n-1) \frac{X_i X_j (1 - X_i - X_j)}{X_i X_j (1 - 2X_i)(1 - 2X_j)} \gamma \quad (5)$$

$$\gamma = \frac{1}{2} + \frac{1}{2} \sum_{i \in U} \frac{1 - 2X_i}{X_i}$$

Il a été démontré que ces π_{ijk} satisfont aux restrictions de deuxième ordre pour π_{ij} :

$$\sum_{i, j \in U (j \neq i)} \pi_{ij} = n(n-1),$$

$$\sum_{j \in U (j \neq i)} \pi_{ij} = (n-1) \pi_i.$$

et

$$\pi_{ijk} = \pi_i \pi_j (c_i + c_j) / 2, \quad (6)$$

et

$$c_i = (n-1)/n\gamma(1-2X_i).$$

Une implication de l'approximation (5) est que $\pi_{ijk}/n(n-1)$ ne dépend pas de n . Par conséquent, l'approximation correspondante de p_z ne dépend pas de n (rappelons que nous avons présupposé que $X_i < 1/n$).

Cette non-dépendance à n serait également survenue si nous avions utilisé l'approximation proposée par Hartley et Rao (1962) pour l'échantillonnage PPT aléatoire :

$$\pi_{ijHR} = n(n-1) X_i X_j$$

$$\{1 + X_i + X_j - \mu_x + 2(X_i' + X_j' + X_i X_j) - 3\mu_x(X_i + X_j - \mu_x - 2\sum_{i \in U} X_i^2)\},$$

(7)

où $\mu_x = \sum_{i \in U} X_i^2$ (rappelons que $\mu_z = \sum_{i \in U} X_i Z_i$). Manifestement, $\pi_{ijHR}/n(n-1)$ ne dépend pas de n . À l'époque, Hartley et Rao ont supposé que $n = O(1)$ lorsque $N \rightarrow \infty$. De plus, en se reportant à une conversation privée avec J.N.K. Rao, Thompson et Wu (2008) déclarent que l'approximation (7) est valide lorsque $n = o(N)$ lorsque $N \rightarrow \infty$. Pour un exemple où (5) et (7) ne peuvent pas être utilisés pour toutes valeurs de n et N , voir l'annexe A.

remplacement en fonction d'un plan d'échantillonnage donné, avec des probabilités d'inclusion du premier ordre de π_i et des probabilités d'inclusion du deuxième ordre de π_{ij} ($i, j = 1, \dots, N$). L'estimateur HT de la population totale, $Y = \sum_{i \in U} Y_i$, est défini par $\hat{Y}_{HT} = \sum_{i \in U} Y_i / \pi_i$. Supposons qu'il y a une mesure de la taille relative X_i (c'est-à-dire $X = \sum_{i \in U} X_i = 1$), de manière à ce que $X_i \leq 1/n$. En fait, on présume ici que les unités où $X_i > 1/n$ sont regroupées dans une strate de certitude distincte. Lorsque les π_i sont proportionnels à ces mesures de la taille, $\pi_i = nX_i$. En définissant que $Z_i = Y_i/X_i$, nous pouvons écrire Y comme moyenne pondérée de Z_i , c'est-à-dire que $Y = \mu_z = \sum_{i \in U} X_i Z_i$. De même, nous pouvons écrire l'estimateur HT de Y dans l'échantillonnage PPT aléatoire comme $\hat{Y}_{HT} = \hat{Y}_{PPT} = \bar{z}$, où \bar{z} est la moyenne de l'échantillon de Z_i .

où $\pi_i = \pi_i$. La première équation est attribuée à Horvitz et Thompson (1952), et la deuxième, à Sen (1953) et à Yates et Grundy (1953). L'expression de rechange suivante pour la variance est mieux adaptée à nos besoins :

$$\text{var}(\hat{Y}_{PPT}) = \text{var}(\bar{z}) = \{1 + (n-1)p_z\} \frac{n}{\sigma_z^2}, \quad (3)$$

$$\text{où } \sigma_z^2 = \sum_{i \in U} X_i (Z_i - \mu_z)^2, \text{ et}$$

$$p_z = \sum_{i \in U} \sum_{j \neq i} \frac{\pi_{ij}}{n(n-1)} \left(\frac{Z_i - \mu_z}{Z_j - \mu_z} \right) \left(\frac{\sigma_z}{\sigma_z} \right). \quad (4)$$

Pour une preuve de (3), voir Knottnerus (2003, page 103). Soulignons que σ_z^2/n aurait été la variance si l'échantillon avait été tiré avec remplacement, avec des probabilités de tirage de X_i .

Le coefficient d'autocorrélation de l'échantillonnage p_z dans l'échantillonnage systématique avec probabilités égales : voir, par exemple, Cochran (1977, pages 209 et 240) et Särndal, Swensson et Wretman (1992, page 79). Soulignons que p_z est un paramètre de population fixe. L'expression *autocorrélation d'échantillonnage* est employée parce que p_z désigne l'autocorrélation entre deux observations sélectionnées au hasard, disons z_{s1} et z_{s2} , à partir de s . Par conséquent, la valeur de p_z dépend du plan d'échantillonnage. Plus précisément, dans le cas d'un échantillonnage avec remplacement, $p_z = 0$, tandis que selon l'EAS, $p_z = -1/(N-1)$.

À propos de l'efficacité de l'échantillonnage à probabilité proportionnelle à la taille aléatoire

Paul KNOTTNERUS¹

Résumé

Dans le présent document, on examine l'efficacité de l'estimateur Horvitz-Thompson au moyen d'un échantillon systématique de probabilité proportionnelle à la taille (PPT) tiré d'une liste en ordre aléatoire. Plus précisément, l'efficacité est comparée avec celle d'un estimateur par quotient ordinaire. Les résultats théoriques sont confirmés d'une manière empirique à l'aide d'une étude de simulation basée sur des données hollandaises de l'Indice des prix à la production.

Mots clés : Estimateur de Horvitz-Thompson ; Indice des prix à la production ; estimateur par quotient ; coefficient d'autocorrélation d'échantillonnage.

1. Introduction

Lorsque la variable y à l'étude dans une population de N unités est plus ou moins proportionnelle à une variable de taille x , on peut s'appuyer sur l'estimateur par quotient d'un échantillon aléatoire simple (EAS) de taille n sans remplacement. Un estimateur de rechange en pareil cas est l'estimateur Horvitz-Thompson (HT), associé à un échantillon systématique à probabilité proportionnelle à la taille tiré d'une liste en ordre aléatoire, ci-après appelé l'échan-

Ces dernières années, plusieurs auteurs ont étudié les procédures d'estimation de la variance pour l'estimateur HT à partir d'un échantillon PPT aléatoire. Voir, entre autres, Brewer et Donadio (2003), Cumberland et Royall (1981), Deville (1999), Knottnerus (2003), Kott (1988 et 2005), Rosén (1997) et Stehman et Overton (1994). Pour une comparaison entre l'efficacité de l'estimateur par quotient et celle de l'estimateur PPT aléatoire, le lecteur est invité à consulter le rapport de Foreman et Brewer (1971), Cochran (1977) et les références fournies dans le présent document. L'un des inconvénients de ces comparaisons est que la correction pour population finie n'est pas prise en compte. Hartley et Rao (1962) tiennent compte de la correction pour population finie, mais sans formule explicite pour l'efficacité. En examinant de plus près les résultats de Gabler (1984), Qualité (2008) démontre que l'estimateur HT connexe d'un échantillonnage réjectif de Poisson de taille n est plus efficace que l'estimateur de Hansen-Hurwitz pour un plan d'échantillonnage avec remplacement. Aucune formule pour l'efficacité accrue n'est toutefois donnée.

Le principal objectif du présent document est d'établir des formules pour l'efficacité de l'estimateur PPT aléatoire par rapport à l'estimateur par quotient. À cette fin, nous présentons une formule simple pour la variation de la taille de l'échantillon requise pour maintenir la même variance lorsqu'un estimateur PPT aléatoire est remplacé par un

estimateur par quotient. De la perspective du plan de sondage, ces formules sont valides lorsque $n = o(N)$, pour $N \rightarrow \infty$. Cette condition suggère que la correction pour population finie peut être omise pour ce genre de plan d'échantillonnage. Fait étonnant, comme nous le verrons dans un exemple à la section 4, l'échantillonnage PPT aléatoire peut réduire la variance de plus de 30 % comparativement à l'échantillonnage PPT avec remplacement, même lorsque la fraction d'échantillonnage n/N est beaucoup plus petite que 30 % ; voir aussi Kott (2005, page 436). De plus, les formules demeurent appropriées d'une perspective basée sur un modèle lorsque n et N sont du même ordre, à condition que N soit grand et que le modèle hypothétique pour les observations Y_i ($i = 1, \dots, N$) satisfasse à des conditions peu contraignantes.

Les grandes lignes du document sont décrites ci-après. La section 2 décrit une expression de rechange pour la variance de l'estimateur HT en fonction du coefficient d'autocorrélation de l'échantillonnage. L'estimateur de variance correspondant pour l'échantillonnage PPT aléatoire est prouvé comme non négatif avec une probabilité 1. La section 3 présente les formules pour l'efficacité de l'estimateur PPT aléatoire par rapport à l'estimateur par quotient pour divers modèles de données souvent rencontrés dans la pratique. La section 4 montre un exemple de données sur l'indice des prix à la production aux Pays-Bas, qui illustre les gains d'efficacité considérables qui peuvent être obtenus en pratique. Un contre-exemple démontre que l'échantillonnage PPT aléatoire n'est pas toujours avantageux. Le document se termine par un sommaire.

2. Une autre expression de la variance pour l'échantillonnage PPT aléatoire

Si l'on suppose une population $U = \{1, \dots, N\}$, et que s est un échantillon de taille fixe n tiré de U sans

$V_{e,o}[d]$

$$= \frac{C_0}{2} \left(c_{11}^2 + \sqrt{(1-p_1)S_2^b - (1-p_{II})\bar{S}_2^w} / M \right) \times \left[(1-p_1)S_2^b \right.$$

$$\left. + \left(\sqrt{\frac{c_{11}^2}{c_{12}^2} \frac{(1-p_1)S_2^b - (1-p_{II})\bar{S}_2^w}{M}} - \frac{1}{M} \right) (1-p_{II})\bar{S}_2^w \right] - \frac{N}{2(1-p_1)S_2^b}$$

$$= \frac{C_0}{2} \left\{ (1-p_1)S_2^b c_{11}^2 \right.$$

$$\left. + (1-p_{II})\bar{S}_2^w \sqrt{c_{12}^2 c_{II}^2 \frac{(1-p_1)S_2^b - (1-p_{II})\bar{S}_2^w}{M}} \right]$$

$$\times \sqrt{\frac{(1-p_{II})\bar{S}_2^w c_{12}^2 c_{II}^2}{(1-p_1)S_2^b - (1-p_{II})\bar{S}_2^w}} / M$$

$$+ (1-p_{II})\bar{S}_2^w \left(c_{II}^2 - \frac{c_{11}^2}{M} \right) \left\{ \right.$$

$$\left. - \frac{N}{2(1-p_1)S_2^b} \right.$$

Preuve de la proposition 7. En ignorant les termes de correction pour population finie d'ordres $O(N^1)$ et $O(M^{-1})$, l'équation (3.3) peut s'écrire :

$$V_{e,i}[d] \approx \frac{C_0}{4(c_1 + \sqrt{c_{II}^2 \bar{S}_2^w / S_2^b})} \left[S_2^b + \left(\sqrt{\frac{c_{II}^2}{c_1} \bar{S}_2^w S_2^b} \right) \right]$$

$$= \frac{C_0}{4} \left(c_1 S_2^b + c_{II} \bar{S}_2^w + 2\sqrt{c_{II}^2 c_{II}^2 \bar{S}_2^w S_2^b} \right)$$

$$= \frac{C_0}{4} \left(\sqrt{c_1 S_2^b} + \sqrt{c_{II} \bar{S}_2^w} \right)^2.$$

De même, l'équation (3.8) peut s'écrire

$$V_{e,o}[d] \approx \frac{C_0}{2} \left[(1-p_1)S_2^b c_{11}^2 \right.$$

$$\left. + 2\sqrt{c_{12}^2 c_{II}^2 (1-p_1)S_2^b (1-p_{II})\bar{S}_2^w} + (1-p_{II})\bar{S}_2^w c_{II}^2 \right] = \frac{C_0}{2} \left[\sqrt{(1-p_1)S_2^b c_{11}^2} + \sqrt{(1-p_{II})\bar{S}_2^w c_{II}^2} \right]^2.$$

L'énoncé de la proposition 7 découle directement de ces deux expressions.

Bibliographie

- Binder, D.A., et Hidiroglou, M.A. (1988). Sampling in time. Dans *Handbook of Statistics*, (Eds, P.R. Krishnaiah et C.R. Rao), North Holland, Amsterdam, 6, 187-211.
- Cochran, W.G. (1977). *Sampling Techniques*, 3^e Ed., New York : John Wiley & Sons, Inc.
- Eckler, A.R. (1955). Rotation sampling. *Annals of Mathematical Statistics*, 26(4), 664-685.
- Ernst, L.R. (1999). The maximization and minimization of sample overlap problems: A half century of results. Rapport technique, U.S. Bureau of Labor Statistics.
- Fuller, W.A. (1999). Environmental surveys over time. *Journal of Agricultural, Biological and Environmental Statistics*, 4(4), 331-345.
- Groves, R.M. (1989). *Survey Errors and Survey Costs*. New York : John Wiley & Sons, Inc.
- Hansen, M., Hurwitz, W.N. et Madow, W.G. (1953). *Sample Survey Methods and Theory*. New York : John Wiley & Sons, Inc.
- Kish, L. (1995). *Survey Sampling*, 3^e Ed., New York : John Wiley & Sons, Inc.
- Lehtonen, R., et Pahkinen, E. (2004). *Practical Methods for Design and Analysis of Complex Surveys*, *Statistics in Practice*, 2^e Ed., New York : John Wiley & Sons, Inc.
- Mas-Colell, A., Whinston, M.D. et Green, J.R. (1995). *Microeconomic Theory*, Oxford University Press, Oxford, UK.
- McDonald, T.L. (2003). Review of environmental monitoring methods: Survey designs. *Environmental Monitoring and Assessment*, 85, 277-292.
- Neyman, J. (1938). Contribution to the theory of sampling human populations. *The Journal of the American Statistical Association*, 33, 101-116.
- Patterson, H.D. (1950). Sampling on successive occasions with partial replacement of units. *Journal of the Royal Statistical Society, Séries B*, 12(2), 241-255.
- Rao, J.N.K., et Graham, J.E. (1964). Rotation designs for sampling on repeated occasions. *Journal of the American Statistical Association*, 59(306), 492-509.
- Särndal, C.-E., Swensson, B. et Wretman, J. (1992). *Model Assisted Survey Sampling*. New York : Springer.
- Scott, C.T. (1998). Sampling methods for estimating change in forest resources. *Ecological Applications*, 8(2), 228-233.
- Thompson, M.E. (1997). *Theory of Sample Surveys. Monographs on Statistics and Applied Probability*, New York : Chapman & Hall/CRC, 74.
- Thompson, S.K. (1992). *Sampling*. New York : John Wiley & Sons, Inc.
- Wolter, K.M. (2007). *Introduction to Variance Estimation*, 2^e Ed., New York : Springer.

La variance de l'estimateur de la différence peut être calculée en utilisant (2.15).

culée en utilisant (2.15).

Sous les conditions de symétrie, $\kappa = 1$, et

$$D = 4[2(1 - p_1)S_2^b - 2\bar{S}_2^w/M]c_{II}c_{I}c_{I2}\bar{S}_2^w$$

est non négative, à moins que l'expression entre les crochets soit négative (ce qui ne peut se produire que si p^1 est grande et que M est petite. Dans ce cas, une solution de coin $m = M$ est réalisée). En outre,

$$\frac{W/\sqrt[2]{S-\frac{q}{2}S(I^d-1)}\sqrt[2]{C_1^{12}C_1^{12}C_1^{12}}}{C_0} = \frac{C_1^{12}C_1^{12}C_1^{12}}{C_0} = u$$

$$[p]^{(\nu, \gamma)} \Lambda$$

$$\frac{C_0}{2}(c_{I12}^2 + 2mc_{II}) \times \left[\frac{N}{2(1-d_I)^2} - \left[\frac{W}{2} \frac{c_{I12}^w}{S_2^w} (d_{II}-1) - \frac{m}{2} \frac{W}{S_2^w} (d_{II}-1) + \frac{m}{2} S_2^2 (d_I-1) \right] \right] = \frac{C_0}{2} \left\{ c_{I12}^2 (d_I-1) + 2 S_2^2 (d_I-1) \right\}$$

$$\frac{N}{\underline{z}_2^q(\mathbf{d}-1)2} - \left\{ \left[\underline{z}_2^u(\mathbf{d}-1) \frac{W}{2} - \underline{z}_2^q(\mathbf{d}-1) \right]_{\parallel} \omega \underline{z}_2 + \underline{z}_2^u(\mathbf{d}-1) \frac{u}{2} + \right.$$

$$= \frac{C^0}{2} \left\{ c_{l_1 l_2}^g(d-1)z + c_{l_1 l_2}^q(d-1)z + \sqrt{\frac{W}{c_{l_1 l_2}^w}} \left[2z - \frac{W}{c_{l_1 l_2}^w} \right] \right.$$

Enfin, la variance de l'estimateur de la différence est

$$\frac{\frac{W/\tau^m \mathcal{S}(\Pi d-1) - \tau^q \mathcal{S}(\Gamma d-1)}{\tau^{\frac{1}{2}} \mathcal{S}(\Pi d-1)} + \tau^{\frac{1}{2}}}{\tau^{\frac{1}{2}} \mathcal{S}(\Pi d-1)} = \frac{w_{\Pi} \tau^{\frac{1}{2}} + \tau^{\frac{1}{2}}}{\tau^{\frac{1}{2}} \mathcal{S}(\Pi d-1)} = u$$

D'après le budget de l'enquête (3.6),

$$\begin{aligned}
& \frac{W_{\tau}^m S_{\parallel}(\mathbf{d}-1) - \frac{q}{z} S_{\parallel}(\mathbf{d}-1) \frac{\tau_1}{\tau_2}}{\frac{\tau_1}{z} S_{\parallel}(\mathbf{d}-1)} \Bigg| = \\
& \frac{\frac{\tau_1}{\tau_2} [W_{\tau}^m S_{\parallel}(\mathbf{d}-1) - \frac{q}{z} S_{\parallel}(\mathbf{d}-1)]}{\frac{\tau_1}{z} S_{\parallel}(\mathbf{d}-1)} \Bigg| = w \\
& \cdot 0 = [\frac{\tau_1}{z} S_{\parallel}(\mathbf{d}-1)] -
\end{aligned}$$

$$\begin{aligned} & \mathcal{Z}[w_{\mathbb{I}}^{\mathcal{Z}} \mathcal{W}_{\mathcal{Z}}^q (\mathbb{I}^{\mathcal{D}} - \mathbb{I}) - \mathcal{Z}_{\mathbb{I}} \mathcal{W}_{\mathcal{Z}}^q S_{\mathbb{I}} (\mathbb{I}^{\mathcal{D}} - \mathbb{I})] \\ \mathbb{I}_0 = & (w_{\mathbb{I}}^{\mathcal{Z}} \mathcal{Z}_{\mathbb{I}} + \mathcal{Z}_{\mathbb{I}} \mathcal{Z})^{\mathbb{I}} \mathcal{W} (\mathbb{I}^{\mathcal{D}} - \mathbb{I}) - \\ & \mathcal{Z}_{\mathbb{I}} \mathcal{W}_{\mathcal{Z}}^{\mathbb{I}} \mathcal{Z}_{\mathbb{I}} (\mathbb{I}^{\mathcal{D}} - \mathbb{I}) (w - \mathcal{W}) + \mathcal{Z}_{\mathbb{I}} \mathcal{Z}_{\mathbb{I}} w \mathcal{W}_{\mathcal{Z}}^q S_{\mathbb{I}} (\mathbb{I}^{\mathcal{D}} - \mathbb{I}) \end{aligned}$$

D'ou,

$$= \frac{m \mathcal{C}_I^{12} \mathcal{C}_I^{12}}{\mathcal{Z}_2^m (\mathcal{D} - 1)} + \frac{(m \mathcal{C}_I^{12} + \mathcal{C}_I^{12}) m}{\mathcal{Z}_2^m (\mathcal{D} - 1)} \left(\frac{\mathcal{M}}{m} - 1 \right) = -\lambda n^2 / 2$$

En exprimant $-\lambda n^2$ d'après ces conditions, nous obtenons :

$${}^0C - mu_{\text{II}}^2C + u_{\text{I}}^2C = \frac{\mathcal{U}}{\mathcal{T}} = 0$$

$$\chi_{12} - \frac{m}{S_2} = \frac{m}{T} = 0$$

$$-2\left(1-\frac{M}{m}\right)\left(1-\rho_{\Pi}^2\right)\frac{m_2}{S_{\Pi}^2}-\lambda(c_{\Pi}c_{12}+c_{12}^{\Pi}m),$$

$$\frac{z^u}{z^q S(d-1)} z^- = \frac{u\varrho}{T\varrho} = 0$$

Les conditions de premier ordre sont :

$$- \lambda (c_1^2 n + c_{12}^2 m - C_0).$$

$$T(n, m, \gamma) = \frac{m}{\gamma S_2^m (d-1)} \left(\frac{W}{m} - 1 \right) \tau_2 + \frac{n}{\gamma S_2^m (d-1)} \left(\frac{N}{n} - 1 \right) \tau_2$$

Preuve de la proposition 6. La fonction lagrangienne de minimisation de (2.15) sous la contrainte (3.6) est

Preuve de la proposition 5. La fonction lagrangienne de minimisation de (2.13) sous la contrainte (3.4) est

minimisation de (2.13) sous la contrainte (3.4) est

$$\begin{aligned} & \left({}^0\mathcal{O} - \tau_{uu} \tau_{\Pi}^{\mathcal{O}} + {}^1\tau_{uu} \mathcal{O} + u {}^1\mathcal{O} \right) \chi - \\ & \frac{\tau_{uu}}{\tau_{\underline{\underline{S}}}} \left(\frac{W}{\tau_{uu}} - 1 \right) + \frac{{}^1\tau_{uu}}{\tau_{\underline{\underline{S}}}} \left(\frac{W}{{}^1\tau_{uu}} - 1 \right) + \\ & \frac{u}{\tau_{\underline{\underline{S}}}({}^1\mathcal{O} - 1)} \left(\frac{N}{u} - 1 \right) \mathcal{Z} = (\chi \tau_{uu} {}^1\tau_{uu} {}^u\tau_{\underline{\underline{S}}}) \mathcal{T} \end{aligned}$$

Les conditions de premier ordre sont :

$$\frac{\partial u}{\partial T} = - \frac{u}{S_2^q} (1 - \rho_1) \frac{u}{z} - \left(\frac{W}{m_1} - 1 \right) \frac{u}{S_2^{1w}} \frac{u}{z}$$

qui peut s'écrire sous la forme

$$\left[z^w \underline{S} \left(\frac{W}{m} - 1 \right) + z^q S^q w \right]_{\Pi} c w - z^w \underline{S} (m_{\Pi} c + c) = 0$$

$$\begin{aligned} & [\frac{w}{z} \underline{S}(u - W) + \frac{q}{z} \underline{S} W]_{\Pi} c u - \frac{w}{z} \underline{S} W (u_{\Pi} c + \cdot^c) = \\ & \frac{w}{z} \underline{S}_{\Pi} c u + \frac{w}{z} \underline{S} W_{\Pi} c u - \frac{q}{z} \underline{S} W_{\Pi} c u - \frac{w}{z} \underline{S} W_{\Pi} c u + \frac{w}{z} \underline{S} W_{\Pi} c u + \frac{w}{z} \underline{S} W_{\Pi} c u = \\ & (\frac{w}{z} \underline{S} W_{\Pi} c u - \frac{q}{z} \underline{S} W_{\Pi} c u)_{\Pi} c u - \frac{w}{z} \underline{S} W_{\Pi} c u + \frac{w}{z} \underline{S} W_{\Pi} c u = \end{aligned}$$

'no, D

$$\frac{W/\tau^u S - \tau^u S}{\tau^u S} \Big|_{\tau^u} = u$$

D'après le budget de l'enquête (3.1), nous trouvons que le

nombre de grappes est

$$\frac{\{z_{1/2}[(M/\Sigma_2^m - \frac{q}{2}S) \Sigma_2^m c_1 c] + z_1 c\} z}{C_0} = \frac{(c_1 m + c) z}{C_0} = u$$

En introduisant ces expressions dans (2.11) et en utilisant les

relations d'égalité (2.9),

$$\frac{um}{S_2^u} \left(\frac{M}{m} - 1 \right) \tau + \frac{u}{S_2^u} \left(\frac{N}{u} - 1 \right) \tau = [p]^{\tau_2} \Lambda$$

$$\frac{C_0}{2(c_1 + m_{\parallel} S_z^g)} \left[\frac{C_0}{2(c_1 + m_{\parallel} S_z^g)} - 1 \right] z =$$

$$+ 4 \left(1 - \frac{M}{m} \right) \frac{m c_0}{S_2^w(c_1 + m c_1)}$$

$$= 2 \left[\frac{C_0}{2(c_I + mc_{II})} - \frac{N}{1} \right] S_2^q$$

$$+ \frac{C_0}{\left(\frac{m}{1} - \frac{M}{1} \right) \left(\frac{m}{1} + \frac{M}{1} \right)}$$

$$= \frac{C_0}{4(c_1 + mc_1)} \left[S_2^b + \left(\frac{m}{1} - \frac{M}{1} \right) \bar{S}_2^w \right] - \frac{N}{2} S_2^q$$

$$4 \left[c_I + \sqrt{c_I c_{II} \underline{S}_2^w / (S_2^w - \underline{S}_2^q)} \right] / M =$$

$$\cdot \frac{qS}{z} \frac{N}{z} - \left[\frac{zS}{z} \left(\frac{W}{1} - \frac{\frac{zS}{z}}{W / \frac{zS}{z} - \frac{qS}{z} \frac{N}{z}} \right) + \frac{qS}{z} \right] \times$$

la U.S. Agency for International Development dans le cadre du Measure Evaluation Project du Carolina Population Center de l'Université de la Caroline du Nord à Chapel Hill, aux termes de l'entente de coopération GPO-A-00-03-00003-00. Les auteurs remercient aussi de leurs commentaires constructifs les participants aux Joint Statistical Meetings (2005) et au XXIII^e Symposium international sur les questions de méthodologie de Statistique Canada (2007).

Annexe

Dans les preuves qui suivent, les espérances, les variances et les covariances sont calculées par rapport au plan de sondage correspondant. Le premier degré de sélection sera désigné par l'indice supérieur I. Le deuxième degré de sélection sera désigné par l'indice supérieur II.

Preuve de la proposition 2. Désignons l'échantillon d'UPPE par S^I , l'échantillon d'USE à la première période par S^{II}_1 , et l'échantillon d'USE à la deuxième période par S^{II}_2 . Alors

$$d = \bar{y}_{2..} - \bar{y}_{1..} = \frac{1}{m} \sum_{i \in S^I_1} \left(\sum_{j \in S^{II}_2} y_{2ij} - \sum_{j \in S^{II}_1} y_{1ij} \right).$$

En désignant les espérances par rapport au premier degré d'échantillonnage par E_I , et celles par rapport au deuxième degré d'échantillonnage par E_{II} , nous avons que la variance sous le plan de d est égale à

$$V[d] = E_I V_{II}[d | S^I] + V_I E_{II}[d | S^I]$$

$$\begin{aligned} &= \frac{1}{I} \left\{ \sum_{i \in S^I_1} V_{II} \left[\sum_{j \in S^{II}_2} y_{2ij} - \sum_{j \in S^{II}_1} y_{1ij} \right] \right\} \\ &+ \frac{1}{I} V_I \left[\sum_{i \in S^I_1} m \bar{K}_{2i} - m \bar{K}_{1i} \right] \\ &= \frac{1}{I} E_I \left\{ \sum_{i \in S^I_1} V_{II} \left[\sum_{j \in S^{II}_2} y_{2ij} - \sum_{j \in S^{II}_1} y_{1ij} \right] \right\} \\ &+ \frac{1}{I} V_I \left[\sum_{i \in S^I_1} m \bar{K}_{2i} - \sum_{i \in S^I_1} m \bar{K}_{1i} \right] \\ &= \frac{1}{I} E_I \left[\sum_{i \in S^I_1} \left(1 - \frac{m}{M} \right) m S^{II}_{2wi} \left(1 - \frac{M}{m} \right) + \left(1 - \frac{m}{M} \right) m S^{II}_{1wi} \right] \\ &+ \frac{1}{I} \left(1 - \frac{n}{N} \right) n (S^{II}_{2b} + S^{II}_{2b} - 2p^I_{1b} S^{II}_{2b}) \\ &+ \frac{1}{I} \left(1 - \frac{n}{N} \right) n (S^{II}_{1b} + S^{II}_{2b} - 2p^I_{1b} S^{II}_{2b}) \\ &= \left(1 - \frac{n}{N} \right) \frac{m}{m} \left(1 - \frac{M}{m} \right) m (S^{II}_{2w} + S^{II}_{1w}) \\ &+ \left(1 - \frac{n}{N} \right) \frac{m}{m} \left(1 - \frac{M}{m} \right) m S^{II}_{1w} \end{aligned}$$

$$\begin{aligned} &= \left(1 - \frac{n}{N} \right) \frac{m}{m} \left(1 - \frac{M}{m} \right) m (S^{II}_{2w} + S^{II}_{1w}) \\ &+ \left(1 - \frac{n}{N} \right) \frac{m}{m} \left(1 - \frac{M}{m} \right) m S^{II}_{1w} \\ &= \left(1 - \frac{n}{N} \right) \frac{m}{m} \left(1 - \frac{M}{m} \right) m (S^{II}_{2w} + S^{II}_{1w}) \\ &+ \left(1 - \frac{n}{N} \right) \frac{m}{m} \left(1 - \frac{M}{m} \right) m S^{II}_{1w} \end{aligned}$$

où la dernière égalité repose sur l'hypothèse des conditions symétriques (2.9).

Preuve de la proposition 3. Désignons l'échantillon d'UPPE par S^I , et l'échantillon d'USE par S^{II} . Alors

$$d = \bar{y}_{2..} - \bar{y}_{1..} = \frac{1}{m} \sum_{i \in S^I} \sum_{j \in S^{II}} (y_{2ij} - y_{1ij}).$$

En désignant les espérances par rapport au premier degré d'échantillonnage par E_I , et celles par rapport au deuxième degré d'échantillonnage par E_{II} , nous avons que la variance sous le plan de d est égale à

$$V[d] = E_I V_{II}[d | S^I] + V_I E_{II}[d | S^I]$$

$$\begin{aligned} &= \frac{1}{I} E_I \left[\sum_{i \in S^I} V_{II} \left(\sum_{j \in S^{II}} (y_{2ij} - y_{1ij}) \right) \right] \\ &+ \frac{1}{I} V_I \left[\sum_{i \in S^I} E_{II} \left(\sum_{j \in S^{II}} (y_{2ij} - y_{1ij}) \right) \right] \\ &= \frac{1}{I} E_I \left[\sum_{i \in S^I} \left(1 - \frac{m}{M} \right) m \left(S^{II}_{2wi} + S^{II}_{2w} - 2S^{II}_{1wi} S^{II}_{2w} \right) \right] \\ &+ \frac{1}{I} V_I \left[\sum_{i \in S^I} m \bar{K}_{2i} - m \bar{K}_{1i} \right] \end{aligned}$$

$$= \frac{1}{I} \left(1 - \frac{m}{M} \right) m \left(S^{II}_{2w} + S^{II}_{2w} - 2p^I_{1w} S^{II}_{2w} \right)$$

$$+ \frac{1}{I} n \left(1 - \frac{N}{n} \right) (S^{II}_{1b} + S^{II}_{2b} - 2p^I_{1b} S^{II}_{2b})$$

$$= \left(1 - \frac{n}{N} \right) \frac{m}{m} \left(1 - \frac{M}{m} \right) m (S^{II}_{2w} + S^{II}_{1w})$$

$$+ \left(1 - \frac{n}{N} \right) \frac{m}{m} \left(1 - \frac{M}{m} \right) m S^{II}_{1w}$$

$$= 2 \left(1 - \frac{n}{N} \right) \frac{m}{m} \left(1 - \frac{M}{m} \right) m$$

$$+ 2 \left(1 - \frac{n}{N} \right) \frac{m}{m} \left(1 - \frac{M}{m} \right) m$$

avec la dernière égalité vérifiée sous les conditions de symétrie

Preuve de la proposition 4. La fonction lagrangienne de minimisation de (2.11) sous la contrainte (3.1) est

$$L(n_1, m_1, n_2, m_2, \lambda) =$$

$$\left(1 - \frac{n_1}{N} \right) \frac{N}{S^2_b} \left(1 - \frac{n_2}{N} \right) \frac{N}{S^2_b} + \left(1 - \frac{n_1}{N} \right) \frac{N}{S^2_b} \left(1 - \frac{n_2}{N} \right) \frac{N}{S^2_b}$$

$$+ \left(1 - \frac{n_1}{N} \right) \frac{N}{S^2_b} \left(1 - \frac{n_2}{N} \right) \frac{N}{S^2_b} + \left(1 - \frac{n_1}{N} \right) \frac{N}{S^2_b} \left(1 - \frac{n_2}{N} \right) \frac{N}{S^2_b}$$

$$- \lambda (c^I_1 n_1 + c^I_{II} n_1 m_1 + c^I_2 n_2 + c^I_{II} n_2 m_2 - C_0).$$

Le cadre multivariées de la section 5 permet d'attribuer des poids d'importance différents aux diverses variances d'intérêt. Des valeurs relativement grandes de α_1, α_2 correspondent à l'importance plus grande des moyennes contemporaines, tandis que de plus grandes valeurs de α_3 correspondent à l'importance plus grande de l'estimation de la variation. Le problème original d'optimisation du plan pour $V[d]$ peut être considéré dans le contexte de (5.1) en posant que $\alpha_1 = \alpha_2 = 0, \alpha_3 = 1$. Ce cadre peut aussi être élargi afin d'inclure les plans destinés à mesurer plusieurs variables. Un défi supplémentaire associé à ce genre de configuration est que les autocorrélations peuvent différer d'une variable à l'autre. Certaines caractéristiques individuelles sont constantes au cours du temps (race, sexe), tandis que d'autres évoluent lentement (logement, dépenses, préférences politiques), et d'autres encore évoluent plus rapidement (revenus ou comportements).

Le présent article traite de trois plans d'échantillonnage et d'un estimateur particulier de la variation, à savoir la différence entre les estimations de la moyenne à deux périodes dans le temps. D'autres options existent en ce qui concerne tant les plans que les estimateurs. Par exemple, dans les plans à échantillon rotatif, une fraction des unités de la première vague est retenue et certaines nouvelles unités sont recrutées. Pour ce genre de plans, l'estimation composée (Hansen et coll. 1953, Patterson 1950, Rao et Graham 1964, Wolter 2007) dans laquelle sont pondérées différemment les contributions des unités indépendantes (celles retirées de l'échantillon après la première vague et celles nouvellement recrutées pour la deuxième vague) et les contributions des unités du panel (utilisées aux deux vagues) produirait des estimations plus efficaces. En général, ce sont des considérations non liées à l'échantillonnage, telles qu'une diminution du fardeau de réponse et la détérioration de la représentativité de l'échantillon en raison de l'évolution de la population, qui motive l'utilisation de ce genre de plans. Ces considérations peuvent être prises en compte dans le modèle de coût (par exemple, un plus grand nombre de rappels nécessaires pour convaincre une unité de répondre) ou dans le modèle de l'erreur totale d'enquête (en introduisant le biais de non-réponse ou de sous-dénombrement et en considérant l'erreur quadratique moyenne d'une estimation plutôt que sa variance sous le plan).

Remerciements

Les auteurs remercient Chris Skinner et John Eltinge de leurs discussions utiles, William Kalsbeek, de ses suggestions aux premières étapes de la rédaction de l'article, ainsi que le rédacteur associé et deux examinateurs de leurs commentaires. Nash Herndon et Oksana Loginova ont amélioré la rédaction. Un appui financier partiel a été fourni par

semblables à celles de nos sections 4 et 5 peuvent encore être obtenues. Si la taille des grappes varie au cours du temps, l'obtention du plan optimal devient une cible mouvante et les plans qui sont optimaux pour les « anciennes » mesures de taille perdront de leur efficacité avec les « nouvelles ».

Dans des ébauches antérieures du présent article, nous avons analysé des plans intermédiaires dans lesquels une fraction non négligeable des unités est retenue et d'autres unités sont échantillonnées indépendamment. Le problème peut alors être considéré comme une minimisation de la variance sous des contraintes d'inégalité appliquées au degré de chevauchement $0 \leq \pi^I \leq 1, 0 \leq \pi^{II} \leq 1$. La théorie générale de l'optimisation sous contraintes non linéaires fait en sorte qu'à condition que la variance de la variation de la moyenne de population D soit monotone en π^I et π^{II} , la solution optimale sera atteinte à l'un des sommets de l'espace des paramètres. Cela justifie notre intérêt pour les trois plans examinés dans le présent article. Ces plans correspondent aux sommets de l'espace des paramètres de coordonnées $(0, 0), (1, 0)$ et $(1, 1)$ pour les plans à échantillons indépendants, à panel de grappes et à panel d'unités d'observation, respectivement. Le point $(0, 1)$ correspond à un plan impossible avec chevauchement complet des unités individuelles sans aucun chevauchement des grappes. Des calculs fastidieux montrent qu'il est possible de satisfaire les conditions de premier ordre dans certains cas intermédiaires également, mais ils correspondent à des maxima locaux de la variance. Bien que ces résultats puissent aussi être intéressants (en ce sens qu'ils fournissent une borne supérieure pour les variances sous le plan), nous ne les avons pas pris en considération dans le présent article. Dans les cas plus compliqués d'optimisation à critères multiples de la section 5, la monotonie n'est pas nécessairement vérifiée et d'autres plans que les trois cas extrêmes considérés dans l'article peuvent produire les valeurs optimales de la fonction objectif (5.2).

Les conditions d'égalité des variances (2.9) peuvent être relâchées au prix de la production d'expressions considérablement plus compliquées. Si les tailles d'échantillon sont maintenues fixes entre les deux vagues de l'enquête, les changements qui suivent peuvent être nécessaires dans toutes les formules pertinentes. Dans les expressions qui ne font pas intervenir les autocorrélations,

$$2S_2^b \mapsto S_2^{lb} + S_2^{2b}, \quad 2S_2^w \mapsto \bar{S}_2^{lw} + \bar{S}_2^{2w} \quad (7.1)$$

tandis que dans les expressions dans lesquelles interviennent les autocorrélations,

$$2(1 - \rho^I)S_2^b \mapsto S_2^{lb} + S_2^{2b} - 2\rho^I S_1^{lb} S_1^{2b}, \quad 2S_2^w(1 - \rho^{II}) \mapsto \bar{S}_2^{lw} + \bar{S}_2^{2w} - 2\rho^{II} \bar{S}_1^{lw} \bar{S}_1^{2w} \quad (7.2)$$

Qualitativement, les résultats seront les mêmes.

grappe dans le cas des grandes grappes. Au départ, nous avons l'intention de considérer des situations dans lesquelles le coût du panel d'USF était plus de deux fois plus élevé que le coût des interviews individuelles. Cependant, comme l'a suggéré l'un des examinateurs, ce coût pourrait être plus faible si les interviews de suivi sont effectuées selon un mode moins coûteux, tel qu'une interview téléphonique ou une enquête par la poste avec questionnaire à remplir soi-même au lieu d'une interview sur place. Le cas échéant, le plan à panel d'unités d'observation semble être le plus rentable des trois.

La structure de population est également sur simplifiée. Nous avons supposé que les grappes sont de taille équilibrée constante. Aucune unité de la population n'en sort et aucune nouvelle unité n'y apparaît. Dans de nombreuses situations pratiques, ces hypothèses sont assez contraignantes. Si la population évolue entre deux vagues de collecte des données, le concepteur d'enquête souhaitera inclure de nouvelles grappes à la deuxième vague, en utilisant les algorithmes de Ernst (1999). Les nouvelles grappes sont placées dans une strate distincte, d'où est tiré un échantillon en grappes. Dans le cas de la NHIS, cette approche est mise en œuvre en servant comme base de sondage d'une liste de « permis de bâtir ». En outre, les effets de mesure dynamiques, tels que le conditionnement et le temps passé dans l'échantillon, donnent lieu à un biais de rotation de l'échantillon, de sorte qu'il pourrait être avantageux de prévoir au moins une certaine rotation des UPE.

En particulier, pour les études réalisées dans le cadre des DHS, le premier argument (couverture) est vraisemblablement plus important que le second (durée de la présence dans l'échantillon), étant donné l'intervalle de temps important entre les cycles de l'enquête (environ cinq ans). La non-réponse et la perte de couverture pourraient sans doute être ajoutées toutes deux au cadre courant comme sources de biais, entraînant ainsi l'optimisation de la moyenne quadratique de l'erreur totale d'enquête plutôt que de la variance sous le plan. Toutefois, la formulation de modèles convaincants de ce genre de biais pourrait être difficile.

Un autre problème qui se poserait si les grappes étaient de tailles différentes est celui de la plus grande gamme de plans applicables. Dans le présent article, nous avons supposé que l'on procédait à un échantillonnage aléatoire simple sans remise aux deux degrés d'échantillonnage. D'autres plans, tels que l'échantillonnage avec probabilités proportionnelles à la taille (PPT), peuvent aussi être utilisés. Pour d'autres plans que l'échantillonnage aléatoire simple, il faudrait se servir de l'estimateur d'Horvitz-Thompson et de l'estimateur de sa variance (Särndal, Swensson et Wretman 1992, Thompson 1997). Les calculs analytiques deviennent compliqués, mais des démonstrations numériques pratiques

de la structure de covariance de la série chronologique d'observations individuelles ainsi que de moyennes de grappes. Il est probable que les résultats seront sensibles au choix d'un modèle partiel. Dans la présente analyse, le problème est plus simple, car il suffit d'avoir un seul paramètre de corrélation pour chaque niveau. Le concepteur d'enquête pourrait devoir introduire plus de paramètres dans le modèle et, conséquemment, étudier la sensibilité du plan choisi à ces divers paramètres.

La complexité du problème, tel qu'il est exposé ci-dessus, peut devenir très rapidement difficile à maîtriser. Nous nous abstentions donc de présenter un traitement plus détaillé dans le présent article.

7. Discussion

Le présent article décrit l'analyse de diverses options de mise en œuvre d'enquêtes à échantillonnage en grappes répétées. Nous fournissons une expression analytique pour les variances sous le plan de l'estimateur par simple différence pour trois plans de sondage fréquemment utilisés (les plans à échantillons indépendants, à panel de grappes et à panel d'unités d'observation). Nous dérivons également les tailles d'échantillon optimales pour l'estimation de la différence entre deux vagues de collecte des données.

Le concepteur d'enquête qui sait que la caractéristique d'intérêt persistera dans une certaine mesure au cours du temps choisira vraisemblablement l'un des plans à collecte par panel, à condition que les coûts des nouvelles visites aux grappes et/ou aux unités d'observation ne soient pas prohibitifs. La comparaison analytique entre les plans à échantillons indépendants et à panel d'unités d'observation est possible et est donnée par la proposition 7. Il convient de préciser que la variance sous le plan de la différence est d'ordre $O(C_0^{-1})$ pour le plan à échantillons indépendants ainsi que pour le plan à panel d'unités d'observation, et qu'elle est d'ordre $O(C_0^{-1/2})$ pour le plan à panel de grappes, où C_0 est le budget total de l'enquête. Le plan à panel de grappes n'est donc viable que pour les petites enquêtes, tandis que le plan à échantillons indépendants ou à panel d'unités d'observation sera vraisemblablement choisi pour les enquêtes à grande échelle.

La structure de coût considérée à la section 3 est assez simpliste. Par exemple, les coûts d'échantillonnage de deuxième degré à la deuxième période pourraient différer selon que les unités sont échantillonnées à partir de nouvelles grappes ou à partir de grappes réunifiées. En outre, le coût peut dépendre de la taille de grappe M_j , car plus de temps et plus de ressources pourraient être nécessaires pour obtenir les cartes et recueillir les données au niveau de la

Si l'enquête à concevoir doit comporter plus de deux vagues de collecte des données, le concepteur de l'enquête pourrait étendre le cadre du problème de maximisation de l'utilité (5.1) en tenant compte des considérations qui suivent.

1. Un plus grand nombre de cibles d'inférence. Les variances possibles que le concepteur d'enquête doit éventuellement prendre en considération peuvent maintenant inclure les variances contemporaines $V[y_1], V[y_2], \dots, V[y_T]$; les différences consécutives $V[y_2 - y_1], \dots, V[y_T - y_{T-1}]$ ou les estimateurs composite/MCG de la variation entre deux périodes adjacentes : d'autres contrastes $V[\sum_i c_i y_i]$.

6. Extension à des vagues multiples

2. La possibilité d'une actualisation. En économie, on spécifie d'habitude les contraintes budgétaires faisant référence à l'avenir sous la forme $\sum_i x_i \delta^i$ où x_i est le montant dépensé au temps t_i et $\delta < 1$ est le facteur d'actualisation associé aux taux d'intérêt. L'actualisation peut également être pertinente pour la fonction d'utilité, et les variances sous le plan qui se situe plus loin dans l'avenir peuvent avoir moins de poids dans le problème d'optimisation.
3. Des formes fonctionnelles inconnues pour les processus chronologiques associés à la variable d'intérêt. Le concepteur d'enquête doit avoir une bonne idée

Figure 12 Variances sous le plan en fonction du budget total C_0 et du coût du panel d'UPE c_{12} . À gauche : lignes de contour de $V_{e,c}$ (pointillé), $V_{e,o}$ (tiret) et $V_{e,l}$ (tiret et point) ; à droite : domaines d'optimalité des trois plans

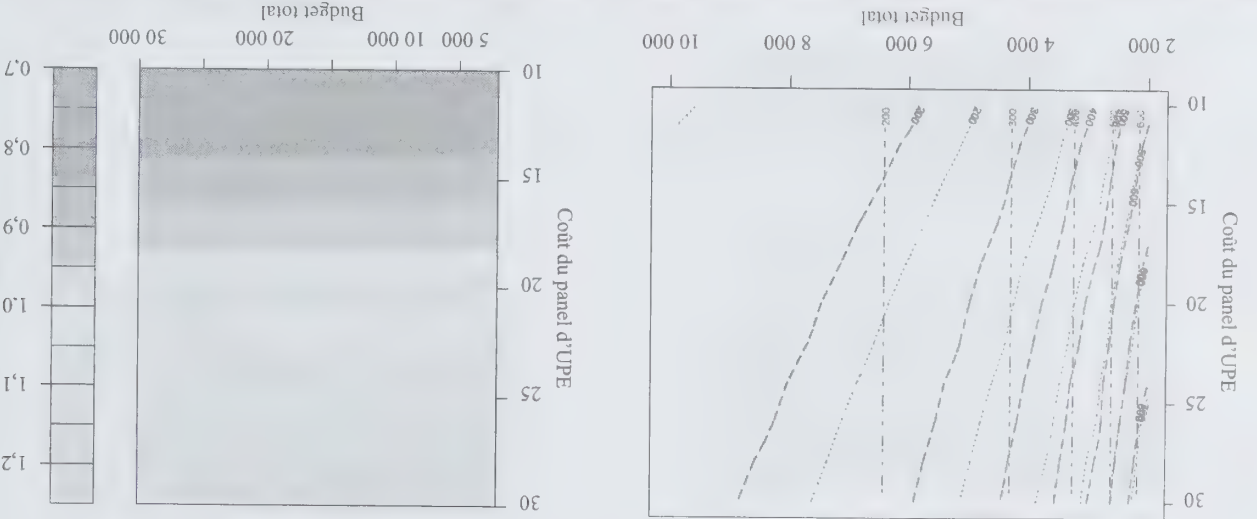
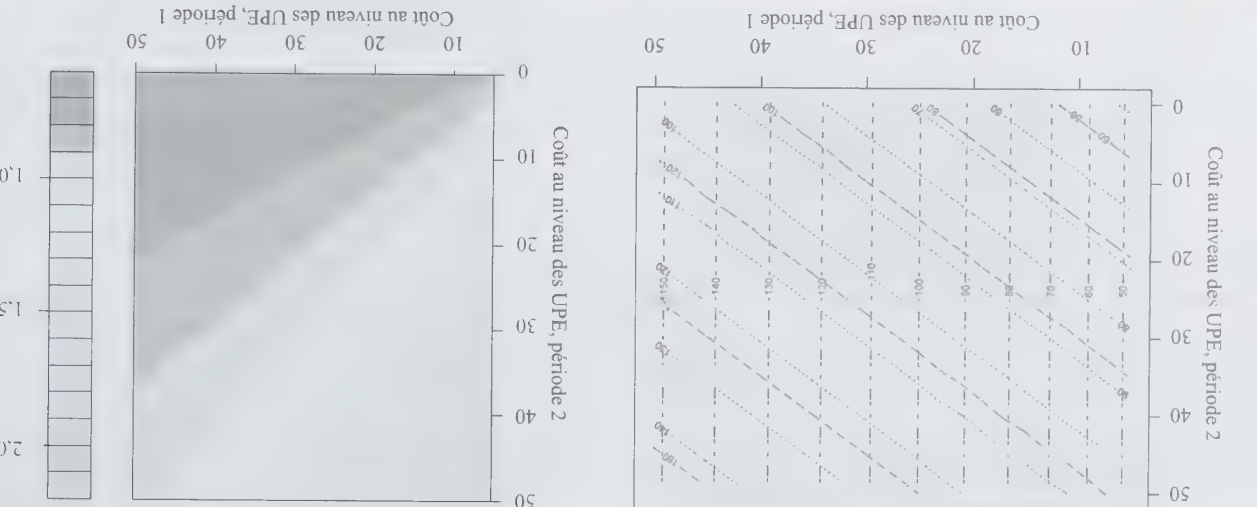


Figure 11 Variances sous le plan en fonction des coûts de collecte des données c_1, c_{12} . À gauche : lignes de contour de $V_{e,c}$ (pointillé), $V_{e,o}$ (tiret) et $V_{e,l}$ (tiret et point) ; à droite : ratio $V_{e,c}/V_{e,l}$



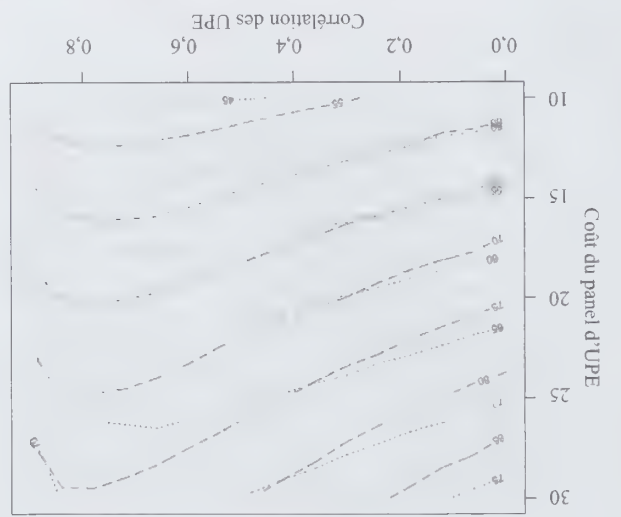


Figure 9 Variances sous le plan en fonction de l'autocorrélation au niveau de la grappe ρ^1 et du coût c_{12}^1 , à gauche : lignes de contour de $V_{e,c}$ (pointillé) et de $V_{e,o}$ (tiret) ; $V_{e,l} = 62,91$; à droite : ratio $V_{e,c}/V_{e,l}$

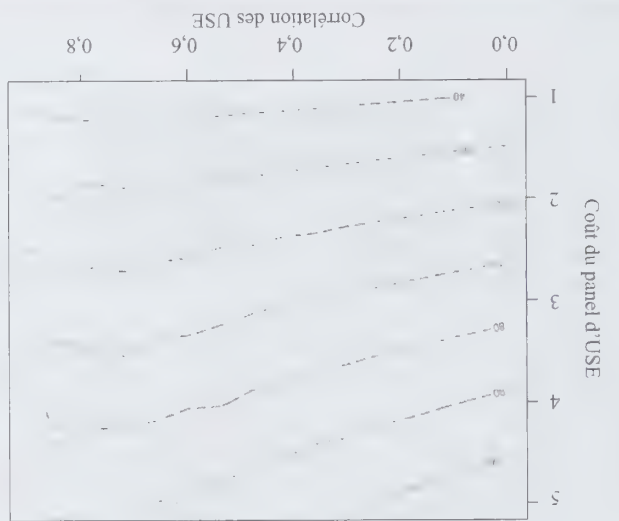
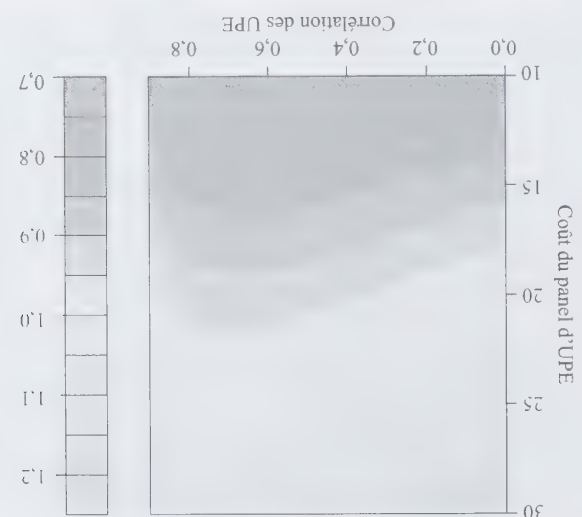
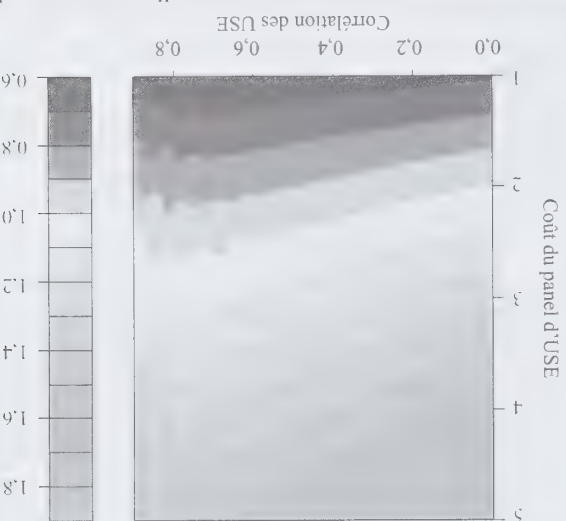


Figure 10 Variances sous le plan en fonction de l'autocorrélation au niveau des unités d'observations ρ^{11} et du coût c_{12}^{11} . À gauche : lignes de contour de $V_{e,o}$ (tiret) ; $V_{e,l} = 62,91$; $V_{e,c} = 59,23$; à droite : ratio $V_{e,o}/V_{e,c}$



La figure 11 est l'équivalent de la figure 5. La partie de gauche montre que le plan à panel d'unités d'observation est moins efficace que le plan à panel de grappes. La partie de droite montre que, si le coût au niveau de la grappe à la deuxième vague dépasse de plus de 15 unités ce coût à la première vague, le plan à échantillons indépendants est plus efficace que le plan à panel de grappes.

Comme nous l'avions conjecturé au début de la présente section, l'intégration des variances des moyennes contenues dans la fonction objectif d'optimisation du plan déplace les préférences du concepteur d'enquête vers des plans plus simples permettant d'échantillonner un plus grand nombre d'unités d'observation finales. Le plan à panel d'unités d'observation ne se justifie maintenant que si les autocorrélations des UPE ainsi que des USE sont élevées, et que les coûts de collecte par panel sont raisonnablement faibles. En outre, le plan à panel de grappes n'est généralement justifié qu'en cas de réduction du coût au niveau de la grappe à la deuxième vague de l'enquête.

Enfin, la figure 12 donne les variances en fonction du budget total de l'enquête et du coût de la collecte de données par panel. Le graphique ne révèle qu'une très légère dépendance à l'égard de C_0 , et le plan à échantillons indépendants est celui qui a la préférence si le mode de collecte par panel est trop coûteux, à savoir si le coût au niveau de la grappe à la deuxième vague dépasse de plus de 107 % le coût à la première vague.

p_{12}^I . Pour des valeurs de $p_{12}^I > 0,7$, la composante $V[d]$ dans (5.2) produit des plans contenant si peu de grappes que cela affecte suffisamment fort $V[p]$ pour nuire à la fonction objectif complète. À cette valeur de l'autocorrélation du panel, le coût de panel maximal auquel le plan à panel de grappes demeure le plus efficace est $c_{12}^I = 24,4$, ce qui signifie que le coût au niveau de la grappe est 44 % plus élevé à la deuxième vague qu'à la première.

La figure 10 montre qu'une plus forte autocorrélation des mesures des USE peut justifier une augmentation modérée du coût associé à la collecte des données. Le coût plus élevé pour lequel le plan à panel d'unités d'observation demeure le plus efficace est $c_{12}^{II} = 2,75$, avec $p_{12}^{II} = 0,78$; cela signifie que le coût de la deuxième interview peut être 75 % plus élevé que celui de la première.

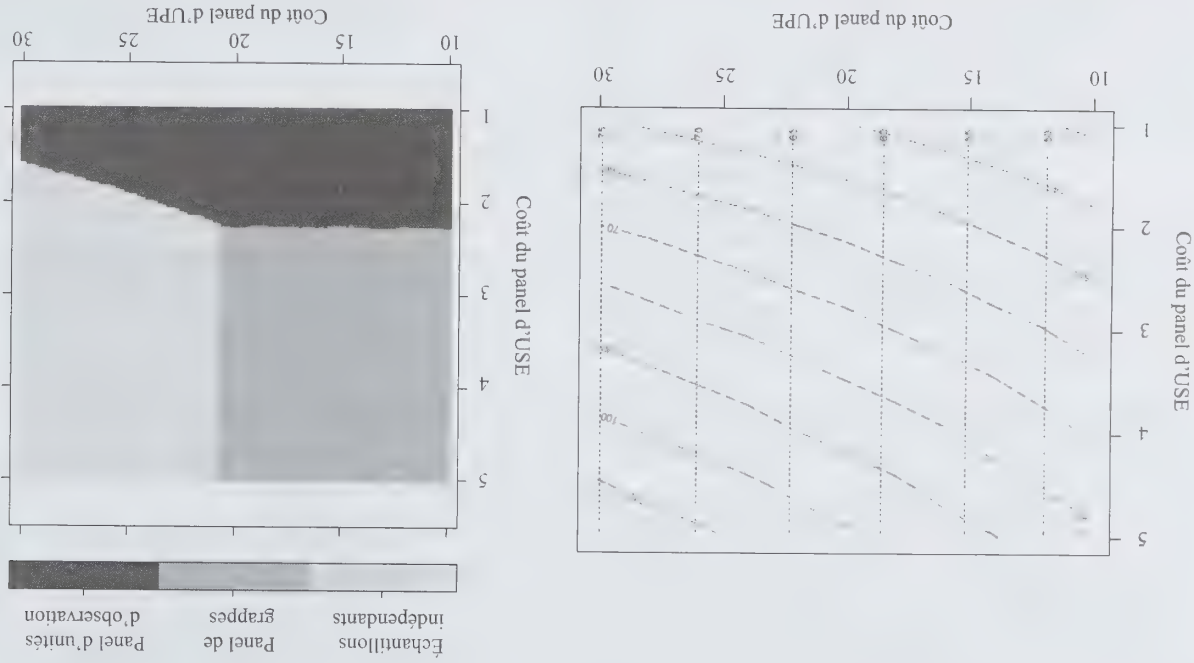


Figure 7 Variances sous le plan en fonction des coûts de collecte des données c_{12}^I, c_{12}^{II} . À gauche : lignes de contour de $V_{e,c}$ (pointillé) et de $V_{e,u}$ (tiret) ; $V_{e,u} = 62,91$, à droite : domaines d'optimalité des trois plans

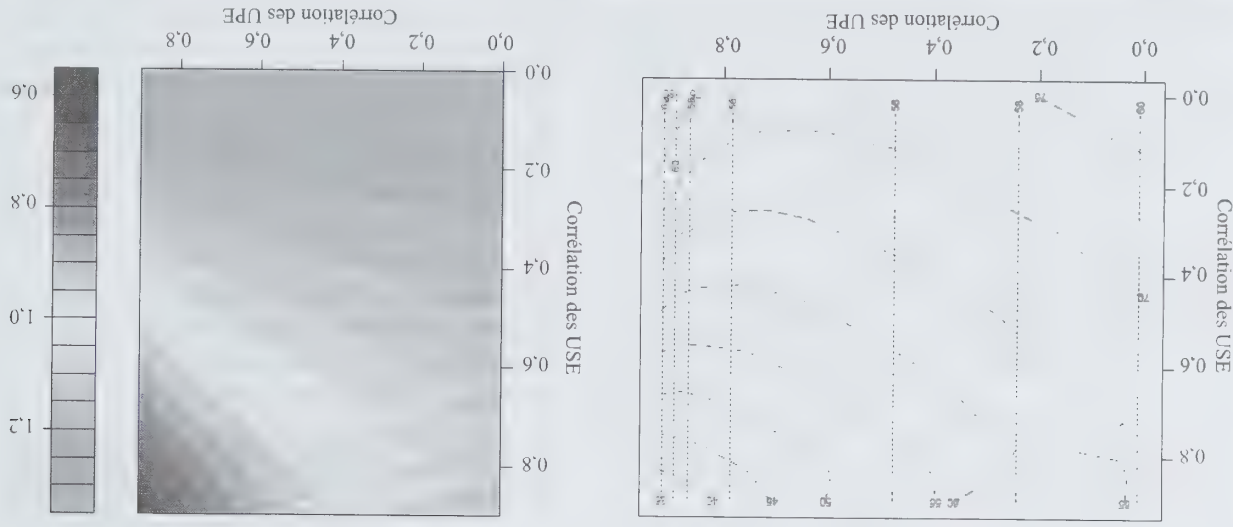


Figure 8 Variances sous le plan en fonction des autocorrélations p_{12}^I, p_{12}^{II} . À gauche : lignes de contour de $V_{e,c}$ (pointillé) et de $V_{e,u}$ (tiret) ; $V_{e,u} = 62,91$, à droite : ratio $V_{e,o}/V_{e,c}$

et d'exprimer le problème d'optimisation comme étant la maximisation de cette expression.

La caractérisation analytique du plan qui optimise (5.2) devient relativement lourde. À la place, nous utilisons l'exemple numérique de la section précédente pour démontrer comment la prise en compte des autres objectifs de conception de l'enquête affecte le choix du plan. Nous devons nous attendre à ce que, dans le cas des plans pour lesquels le coût du suivi est plus élevé ($c_1^I \geq c_1^I + c_2^I$, $c_2^{II} \geq c_2^{II} + c_2^{II}$), les plans plus simples soient sélectionnés plus fréquemment : le plan à panel de grappes peut être préféré au plan à panel d'unités d'observation, et le plan à panel d'échantillons indépendants peut être préféré au plan à panel de grappes. Pour les conditions de base (4.1), nous avons

$$V_{e,1}[\bar{y}] = 49,93, V_{e,1}[\bar{y}] = 47,68, V_{e,1}[\bar{y}] = 61,69, \\ V_{e,1} = 62,91, V_{e,1} = 59,23, V_{e,1} = 70,02.$$

où les indices temporels de $y_{t..}$ sont omis. Le plan à panel d'unités d'observation est assez inefficace pour l'estimation de la moyenne propre à la période, car un plus petit nombre d'unités sont échantillonnées. En revanche, le plan à panel de grappes est le plus efficace, suivi de près par le plan à échantillons indépendants.

Les figures 7 à 12 sont la réplique des figures 1 à 6, respectivement. Puisque le meilleur plan en ce qui concerne V est maintenant le plan à panel de grappes, la plupart de ces graphiques indiquent la préférence pour ce plan. La figure 7 montre que, si l'on tient compte des variances des moyennes contemporaines, les plans à échantillons indépendants et à panel de grappes plus simple sont préférés pour une fraction plus importante de configuration des paramètres, et occupent une part plus grande du graphique dans la figure 1. Le point où les trois plans sont équivalents correspond à $c_1^I = 20,6$, $c_2^{II} = 2,27$, qui est plus proche de l'origine qu'à la figure 1, dans laquelle seule la variance de la différence était prise en compte.

La figure 8 révèle que le plan à panel d'unités d'observation n'est justifié que si les autocorrélations sont toutes deux supérieures à 0,6 (pour les valeurs données des variances de population et des coûts). Rappelons qu'à la figure 2, le plan à panel d'unités d'observation avait la préférence quand $\rho^{II} > 0,34$, le choix étant peu dépendant de ρ^I .

La figure 9 montre comment les corrélations et les coûts au niveau des UPE influencent le choix du plan. Le plan à panel d'unités d'observation est moins efficace que celui à panel de grappes pour toutes les combinaisons de paramètres dans ce graphique. Donc, le choix se fait entre le plan à échantillons indépendants et le plan à panel de grappes. Évidemment, si la collecte des données en mode de panel est coûteuse, le plan à échantillons indépendants sera préféré au plan à panel de grappes. Fait intéressant, la préférence pour un plan particulier n'est pas monotone en

des données sur plusieurs caractéristiques et de nombreux utilisateurs s'intéressent à la production d'estimations contemporaines. Afin de tenir compte des contraintes d'exactitude associées à ces diverses variables et estimations, le concepteur d'enquête doit avoir à l'esprit plusieurs variances quand il choisit le plan à mettre en œuvre. Il s'agit d'un problème d'optimisation à critères multiples et aucun plan individuel ne sera le meilleur pour tous les problèmes d'estimation possibles. Dans le présent contexte, le plan à panel d'unités d'observation peut donner de bonnes estimations de la variation quand l'autocorrélation des UPE ainsi que celle des UPE est élevée, mais risque de produire une petite taille d'échantillon si le coût du suivi des UPE ainsi que des UPE est élevé. Des estimations d'une grande précision pour n'importe quelle période pourraient être obtenues en passant au plan à panel de grappes, voire même au plan à échantillons indépendants.

Dans cette situation, il est possible de comparer les différents plans en utilisant l'argument microéconomique classique de maximisation de l'utilité sous les contraintes budgétaires (Mas-Colell, Whinston et Green 1995). Dans le contexte des plans de sondage, l'utilité pour le concepteur d'enquête augmente avec la précision des estimations d'après les données de l'enquête, ou ce qui est équivalent, diminue quand les variances sous le plan augmentent. Une forme fonctionnelle simple est donnée par la fonction d'utilité de Cobb-Douglas :

$$U(\text{plan}) = V_{-a_1}^{\text{plan}}[\bar{y}_1..] V_{-a_2}^{\text{plan}}[\bar{y}_2..] V_{-a_3}^{\text{plan}}[d], \quad (5.1)$$

Ici, α_1 , α_2 et α_3 sont des constantes positives qui découlent les poids relatifs des trois variances sous le plan dans le processus décisionnel. Dans (5.2), les variances $V[\bar{y}_1]$ et $V[\bar{y}_2]$ sont les variances des moyennes dans les enquêtes avec échantillonnage en grappes données par (2.8). La variance de l'estimateur de la différence correspond à (2.10), (2.12) ou (2.14), selon le plan choisi. Le problème que doit alors résoudre le concepteur d'enquête consiste à maximiser (5.1) sous les contraintes budgétaires particulières au plan (3.1), (3.4) ou (3.6). La maximisation est effectuée sur les paramètres du plan (mode de collecte des données, nombre de grappes à chaque période, nombre d'unités d'observation à chaque période), sachant les caractéristiques de la population (variances et autocorrélations) et le processus de collecte des données (coûts).

Supposons que la précision des trois estimations \bar{y}_1 , \bar{y}_2 et d ont chacune la même importance pour le décideur, de sorte que $\alpha_1 = \alpha_2 = \alpha_3$. Afin d'obtenir une fonction objectif mesurée dans les unités de variance et sur la même échelle que les variances, il est commode de définir une variance multivariée

$$V^{\text{plan}} = (V_{-a_1}^{\text{plan}}[\bar{y}_1..] V_{-a_2}^{\text{plan}}[\bar{y}_2..] V_{-a_3}^{\text{plan}}[d])^{1/3}, \quad (5.2)$$

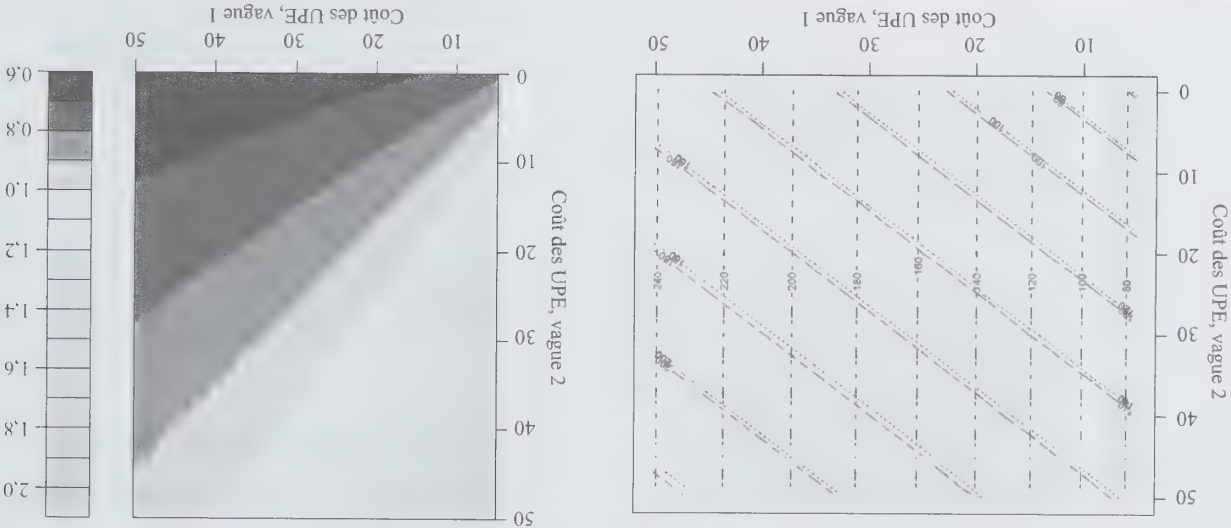


Figure 5 Variances sous le plan en fonction des coûts au niveau de la grappe à la première vague, c_1 , et à la deuxième vague, c_2 . À gauche : lignes de contour de $V_{c_1, c_2}[d]$ (pointillé), $V_{c_1, c_2}[d]$ (tiret) et $V_{c_1, c_2}[d]$ (tiret et point) ; à droite : ratio $V_{c_1, c_2}[d] / V_{c_1, c_2}[d]$

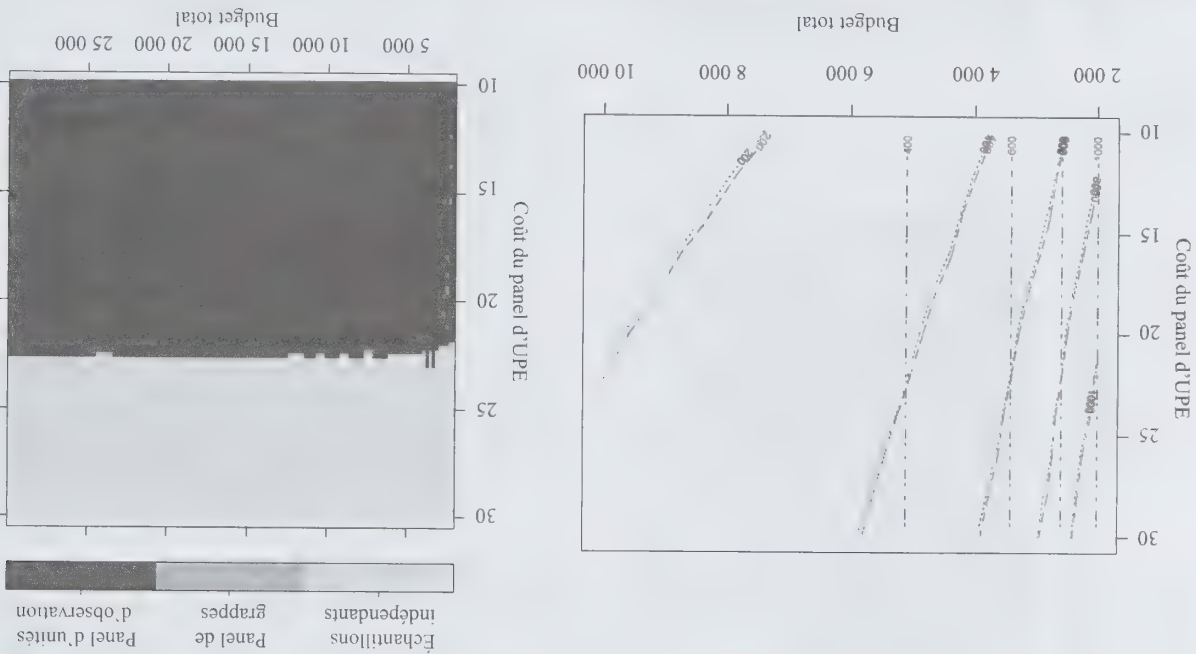


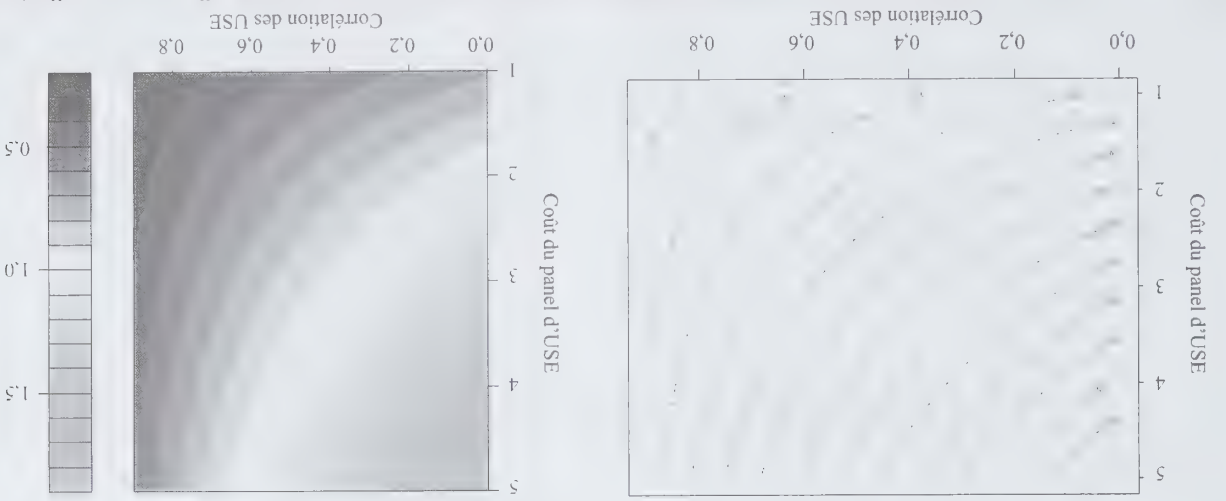
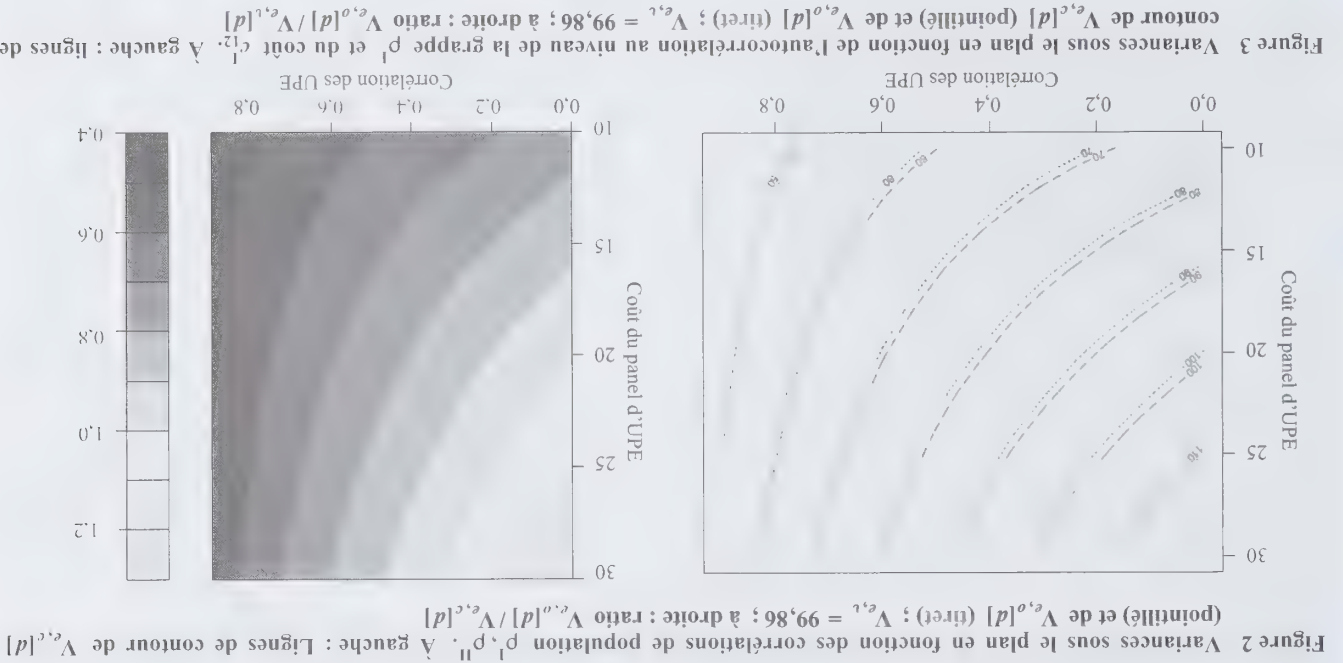
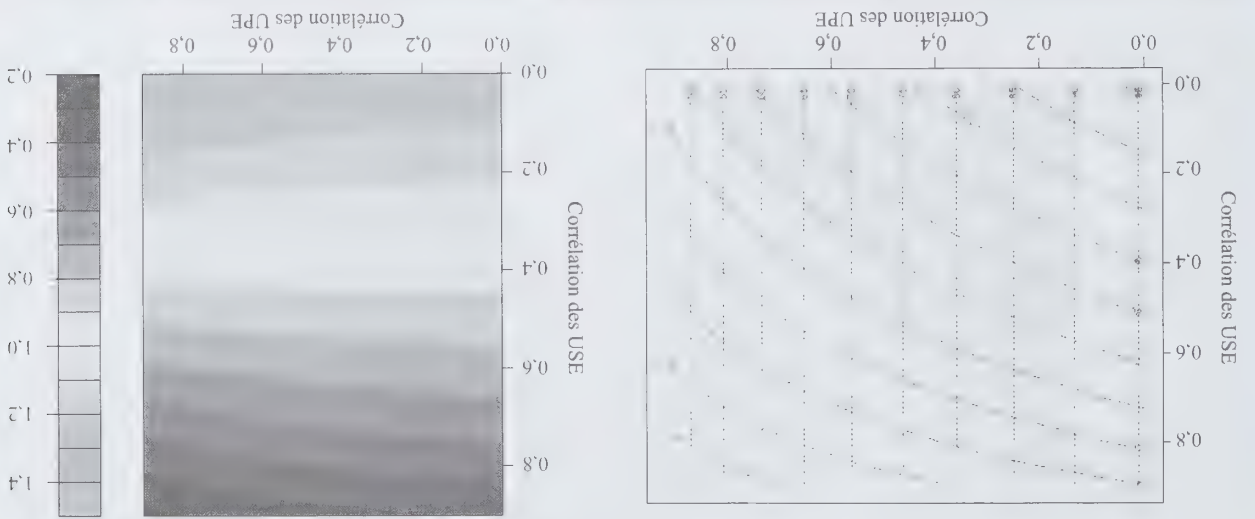
Figure 6 Variances sous le plan en fonction du budget total C_0 et du coût du panel d'UPE, c_{12} . À gauche : lignes de contour de $V_{c_1, c_2}[d]$ (pointillé), $V_{c_1, c_2}[d]$ (tiret) et $V_{c_1, c_2}[d]$ (tiret et point) ; à droite : domaines d'optimalité des trois plans

pourrait être nécessaire pour toute enquête particulière qu'un statisticien doit concevoir.

5. Plan de sondage avec critères multiples

Jusqu'à présent, nous avons limité notre analyse à l'estimation de la différence entre les moyennes d'une seule variable à deux vagues de collecte des données. La plupart des enquêtes à grande échelle sont conçues pour recueillir

Dans l'ensemble, cet exemple numérique montre que, selon les paramètres de la population et les coûts de la collecte des données, chacun des trois plans peut être le plus efficace. De faibles corrélations et des coûts élevés à la deuxième vague ont tendance à favoriser le plan à échantillons indépendants. Vu que les six paramètres de population initiaux et les cinq paramètres de coût pourraient ne pas être représentatifs d'un grand nombre d'enquêtes répétées, une analyse de sensibilité semblable à celle présentée ici



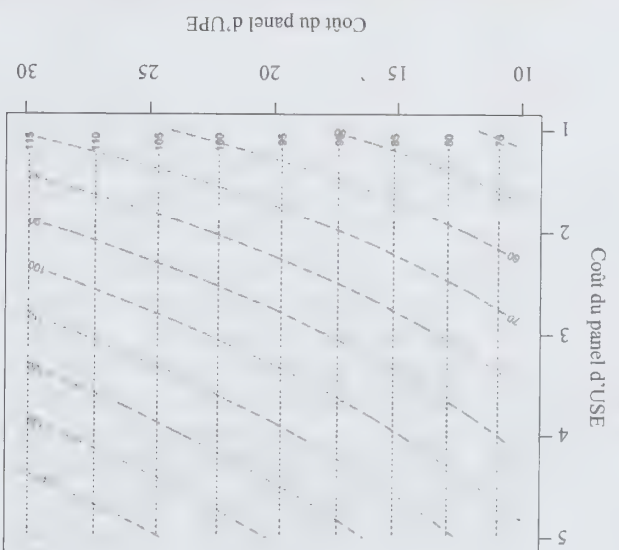
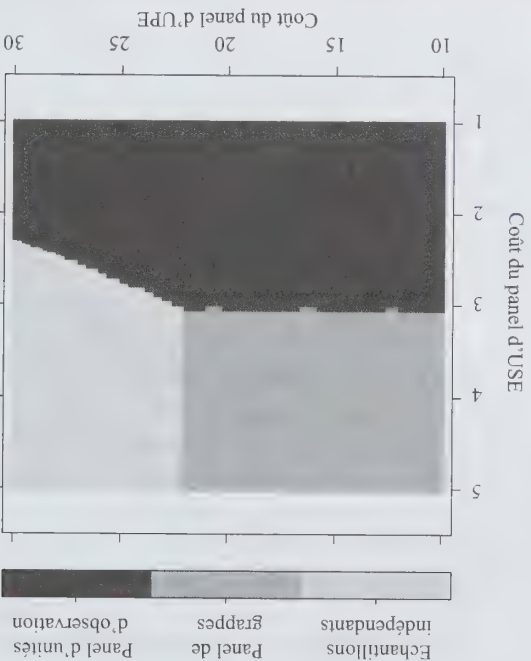


Figure 1 Variances sous le plan en fonction des coûts de collecte des données c_{12}^I, c_{12}^{II} . À gauche : lignes de contour de $V_{e,c}[d]$ (pointillé) et $V_{e,0}[d]$ (tiret) ; $V_{e,L} = 99,86$; à droite : domaines d'optimalité des trois plans



À la figure 3, nous examinons l'effet du coût au niveau de la grappe et de l'autocorrélation sur le choix du plan. Les combinaisons d'un coût élevé de collecte des données à la deuxième vague et d'une faible autocorrélation des UPE, qui figurent dans le coin supérieur gauche du graphique, font du plan à échantillons indépendants celui qui est le plus intéressant. Autrement, le plan à panel d'unités d'observation est celui qui convient le mieux. Notons que les lignes de contour pour les plans à panel d'unités d'observation sont très proches les unes des autres et que les différences de variance entre les deux plans sont inférieures à 2 % dans l'espace entier des paramètres de ce graphique.

À la figure 4, nous examinons l'effet du coût au niveau de l'unité d'observation et de l'autocorrélation sur le choix du plan. Ni la variance sous le plan à échantillons indépendants ni celle sous le plan à panel de grappes n'est affectée par la variation des paramètres présentes dans ce graphique. La variance sous le plan à échantillons indépendants est égale à 99,86, tandis que celle sous le plan à panel de grappes est de 91,37, de sorte que nous comparons le plan à panel d'unités d'observation à ce dernier seulement. Une forte autocorrélation ($\rho^{II} \geq 0,6$) peut justifier le coût très élevé de la deuxième interview (jusqu'à quatre fois plus élevé que celui de la première), mais dans le coin supérieur gauche du graphique, qui correspond aux faibles autocorrélations et aux coûts de panel élevés, le plan à panel de grappes est celui qui donne les meilleurs résultats.

La figure 5 relie les variances sous le plan aux coûts d'enquête au niveau de la grappe. L'axe horizontal donne le meilleur résultat. La figure 6 illustre la relation entre le plan le plus efficace et le budget total de l'enquête, ainsi que le coût du mode panel de collecte des données au niveau de la grappe. Pour $C_0 > 10\,000$, le plan à panel d'unités d'observation est le meilleur si $c_{12}^I < 22,7$, c'est-à-dire si le coût additionnel du mode de collecte par panel des données au niveau de la grappe n'excède pas 127 % du coût initial de collecte au niveau de la grappe à la première vague. Curieusement, pour certaines configurations isolées des paramètres dans de petites enquêtes, le plan à panel de grappes peut donner de meilleurs résultats que le plan à panel d'unités d'observation qui domine le reste du graphique. La différence entre les variances sous le plan à panel de grappes et le plan à panel d'unités d'observation est inférieure à 4 % pour toutes les combinaisons de paramètres dans ce graphique.

4. Exemple numérique

Afin d'illustrer comment les caractéristiques de la population (variances et autocorrélations) et le processus de collecte des données (coûts) influent sur le choix du plan le plus efficace, nous considérons un exemple numérique. Choisissons la configuration de base avec des conditions symétriques et posons que les valeurs des paramètres sont :

$$\begin{aligned} N &= 10\,000, \quad M = 1\,000, \quad S_y^2 = 100, \\ S_w^2 &= 400, \quad \rho^I = 0,1, \quad \rho^{II} = 0,35, \\ c_{II}^I &= c_{II}^{II} = 1, \quad c_{II}^{I2} = 3, \quad c_I^I = c_I^{II} = 10, \\ c_{II}^{I2} &= 18, \quad C_0 = 20\,000. \end{aligned} \tag{4.1}$$

La structure de coût implique que le coût de la collecte de l'information initiale pour une grappe est le coût de dix interviews, tandis que le coût du suivi de la même grappe est celui de huit interviews seulement. Par ailleurs, obtenir la deuxième interview auprès de la même unité coûte deux fois plus cher qu'obtenir la première.

Étant donné ces paramètres, les tailles d'échantillon et les variances sous le plan sont :

$$\begin{aligned} m_{e,c} &= 12, & m_{e,c} &= 12, & m_{e,c} &= 8, \\ n_{e,c} &= 455, & n_{e,c} &= 476, & n_{e,c} &= 476, \\ m_{e,c}n_{e,c} &= 5\,460, & m_{e,c}n_{e,c} &= 5\,712, & m_{e,c}n_{e,c} &= 3\,808, \end{aligned}$$

$$V_{c,I}[d] = 99,86, \quad V_{c,II}[d] = 91,37, \quad V_{c,III}[d] = 90,20. \tag{4.2}$$

Le plan à panel d'unités d'observation est 1,2 % plus efficace que le plan à panel de grappes, et 10,7 % plus efficace que le plan à échantillons indépendants. Cependant, il produit une taille d'échantillon appréciablement plus petite, égale aux deux tiers seulement de celle de l'échantillon sous le plan à panel de grappes et à 70 % de la taille de l'échantillon sous le plan à échantillons indépendants.

Naturellement, ces résultats sont fortement liés aux paramètres de la population et à la structure de coût. Pour nous décrire des profils généraux de changement des variances et, donc, de l'efficacité relative des divers plans quand ces paramètres varient ? Dans (4.2), les variances sont dérivées de 13 paramètres donnés dans (4.1), et il est difficile de formuler des énoncés pertinents au sujet de tous ces paramètres simultanément. Plus bas, nous essaierons de donner des coupes transversales bidimensionnelles de cet espace à 13 dimensions et de présenter des illustrations graphiques de la variabilité des variances sous le plan, et donc les domaines d'optimalité de chaque plan, en faisant varier deux paramètres à la fois. Nous fournissons les graphiques des variances pour les plans concernés (habituellement, le plan à panel de grappes est représenté par des courbes en pointillé ; le plan à panel d'unités d'observation, par des courbes en tirets et le plan à échantillons indépendants, par des courbes en tirets et points. Dans la plupart des

graphiques, le plan indépendant n'est pas affecté par les variations des paramètres qui définissent les axes des graphiques et est donc ignoré). Nous montrons également l'efficacité relative des divers plans en représentant les domaines de l'espace des paramètres en jaune/gris clair si le plan à échantillons indépendants est le plus efficace, en vert/gris moyen si le plan à panel de grappes est le plus efficace, et en violet/gris foncé si le plan à panel d'unités d'observation est le plus efficace (le code R utilisé pour générer les graphiques est disponible au <http://web.missouri.edu/~kolenikov/SMJ2011/>).

La figure 1 illustre comment les variances sous le plan, et donc le plan le plus efficace, varient en fonction des coûts des panels d'UPF et d'USE, c_I^I et c_{II}^{I2} . Manifestement, ces variations n'ont pas d'incidence sur la variance du plan à échantillons indépendants, qui sert de référence. En outre, les variations de c_{II}^{II} n'influencent pas la performance du plan à panel de grappes, qui correspond aux droites d'isovariance verticales en pointillé dans le graphique de gauche. Les courbes descendantes en tirets sont les courbes d'isovariance pour le plan à panel d'unités d'observation. Il convient de souligner que le coin inférieur gauche du graphique correspond à la situation de gratuité dans laquelle la deuxième vague de collecte des données ne coûte rien : les coûts de panel sont égaux aux coûts de la période unique, $c_{II}^{I2} = c_I^I$, $c_{II}^{II} = c_{II}^{I2}$. Quand les coûts de la collecte des données de panel sont extrêmement élevés (coin supérieur droit du graphique), le plan à échantillons indépendants est le plus efficace. Le point où les trois plans ont la même variance correspond à $c_I^I = 22$, $c_{II}^{I2} = 3,05$; autrement dit, le coût de la deuxième interview est 2,05 fois plus élevé que celui de la première, et les coûts au niveau de la grappe sont 20 % plus élevés à la deuxième vague qu'à la première. Néanmoins, une autocorrélation positive justifie la réduction de la taille de l'échantillon du plan à panel d'unités d'observation comparativement au plan à échantillons indépendants. Si le coût du panel au niveau de la grappe est plus faible, mais que le coût de la deuxième interview est plus élevé, le plan à panel de grappes est le plus efficace. Si les deuxième interviews sont plus coûteuses, le plan le plus efficace est le plan à panel d'unités d'observation. Ce dernier domaine comprend notre cas de référence avec $c_{II}^{I2} = 18$ et $c_{II}^{II} = 3$.

La figure 2 montre les variations des variances sous le plan associées aux variations des autocorrélations ρ^I , ρ^{II} . La variance sous le plan à échantillons indépendants n'est pas affectée par cette variation, et celle sous le plan à panel de grappes n'est pas affectée par les variations de ρ^{II} . Le plan à panel d'unités d'observation est plus efficace quand l'autocorrélation des USE est plus forte, $\rho^{II} > 0,34$. Sinon, le plan à panel de grappes donne une variance plus faible.

Le premier terme est le coût au niveau de la grappe et le deuxième terme, le coût des interviews individuelles.

Proposition 6. Pour le plan à panel d'unités d'observation, les tailles d'échantillon optimales sont données par

$$m = \frac{\sqrt{\frac{C_1^2}{(1-p_{II})S_2^2} - (1-p_{II})S_2^2} / M}{C_0} \quad n = \frac{\sqrt{\frac{C_1^2}{(1-p_{II})S_2^2} - (1-p_{II})S_2^2} / M}{C_0} \quad (3.7)$$

La variance sous le plan de l'estimateur de la différence résultant est

$$V_{e,c}[d] = \frac{C_0}{2} \left\{ (1-p_I)S_2^2 C_1^2 + \left[(1-p_I)S_2^2 - \frac{M}{1-p_I} \right] \frac{(1-p_{II})S_2^2 C_1^2}{(1-p_{II})S_2^2 - (1-p_{II})S_2^2 / M} \right. \\ \left. + (1-p_{II})S_2^2 \frac{C_1^2}{(1-p_{II})S_2^2 - (1-p_{II})S_2^2 / M} \right\} - \frac{M}{2(1-p_I)S_2^2} \quad (3.8)$$

Les expressions des tailles d'échantillon (3.7) ressemblent à celles données à l'équation (3.3) pour le plan à échantillons indépendants, avec le coût de la collecte des données dans une seule vague remplacé par le coût de la collecte de données de panel, et les composantes de la variance S_2^2 et S_2^w remplacées par $(1-p_I)S_2^2$ et $(1-p_{II})S_2^w$. La taille de l'échantillon de deuxième degré m dépend uniquement du coût relatif aux niveaux de la grappe et de l'unité d'observation et du ratio des composantes de la variance augmentées des autocorrélations. Donc, comme dans le cas du plan à échantillons indépendants, la relation entre la taille de l'échantillon et la portée de l'enquête n'a lieu que par la voie de $n \propto C_0$, et la variance de la différence diminue de manière inversement proportionnelle à C_0 .

En étendant les relations entre les formes fonctionnelles des équations (3.3) et (3.8), nous pouvons établir les relations générales entre les deux plans :

Proposition 7. Si $M \gg 1$ et $N \gg 1$, alors $V_{e,c}[d] \geq V_{e,c}[d]$ si

$$2 \left(\sqrt{\frac{C_1^2}{(1-p_I)S_2^2} + \sqrt{\frac{C_1^2}{(1-p_{II})S_2^w}}} \right) \geq \left[\sqrt{\frac{C_1^2}{(1-p_I)S_2^2} + \sqrt{\frac{C_1^2}{(1-p_{II})S_2^w}}} \right]^2 \quad (3.9)$$

Malheureusement, la variance sous le plan à panel de grappes que l'on peut obtenir en combinant les résultats de la proposition 5 avec (2.13) ne permet pas une comparaison aussi claire.

$$D = \left(1 + \frac{S_2^2}{S_2^w C_0} \right) + 8 \frac{(1-p_I)S_2^2 C_0}{S_2^w C_0} - 4 \frac{C_1^2}{C_0} \left(1 + \frac{S_2^2}{S_2^w C_0} \right) \geq 0.$$

La variance de l'estimateur de la différence est obtenue en entrant ces expressions dans (2.13). Sous les hypothèses de conditions symétriques aux deux vagues de l'enquête

$$D = 4 - 8 \frac{M C_0}{(1-p_I)S_2^2 C_0} + 8 \frac{M C_0}{S_2^w C_0}, \quad m_1 = m_2 = m$$

$$n = \frac{C_0}{C_1^2 + 2C_0 m} = \frac{C_0}{C_0 + \sqrt{(C_0)^2 - \frac{M}{2(1-p_I)S_2^2 C_0} + \frac{M}{2(1-p_I)S_2^2 C_0}}} = \frac{C_0}{C_0 + 2C_0 \left[1 + \sqrt{1 - \frac{M C_0}{2(1-p_I)S_2^2 C_0} + \frac{M C_0}{2(1-p_I)S_2^2 C_0}} \right]}$$

et $V_{e,c}[d]$ peut être calculée à partir de (2.13).

Fait intéressant, le nombre d'USE dépend des coûts au niveau des USE C_{II} , mais non des coûts au niveau des UPE C_1^2 . Une augmentation de la corrélation intragappe, ou de S_2^b , ou une diminution de S_2^w , entraîne comme cela est à

prévoir une diminution du nombre optimal d'USE et une augmentation du nombre optimal d'UPE. La dépendance des paramètres du plan à l'égard du budget de l'enquête C_0 n'est pas négligeable. Pour de très petites enquêtes, le nombre d'unités par grappe est proportionnel à C_0 , tandis que le nombre de grappes n'est pas affecté par C_0 . En effet, si la caractéristique étudiée présente une forte corrélation entre les périodes d'observation, il est préférable d'obtenir des estimations exactes des moyennes de grappe, desquelles découlerait une bonne exactitude de l'estimateur de la différence globale. Autrement dit, le premier terme de (2.13) est relativement petit en raison du coefficient de corrélation positif p_I , et le deuxième terme est inversement proportionnel à C_0 . Pour les grandes enquêtes, $D \propto C_0$, de sorte que le nombre d'unités par grappe ainsi que le nombre de grappes sont proportionnels à $\sqrt{C_0}$. Le premier terme de (2.13) est alors inversement proportionnel à $\sqrt{C_0}$, et le deuxième terme est inversement proportionnel à C_0 . Une augmentation du budget de l'enquête affectera tous les termes, quoique dans une mesure différente.

3.4 Plan à panel d'unités d'observation

Pour le plan à panel d'unités d'observation, la contrainte budgétaire est donnée par

$$C_0 = C_1^2 n + C_{II}^2 m. \quad (3.6)$$

- c_{12}^{II} est le coût au niveau individuel si l'unité est observée aux deux périodes dans le plan à panel d'unités d'observation (coût du panel d'USE) ;
- C_0 est le budget total alloué au travail sur le terrain aux deux périodes.

Les indices supérieurs en chiffres romains désignent le degré d'échantillonnage. Les indices inférieurs en caractères arabes correspondent à la vague à laquelle l'échantillon est tiré. Les coûts au niveau de la grappe comprennent le coût de l'échantillonnage des grappes, l'obtention des cartes des UPE, la collecte de données sur la collectivité, la formation locale des intervieweurs, etc. Les coûts au niveau individuel sont principalement ceux des interviews sur place auprès des unités d'observation finales. Le coût total C_0 est conçu comme le coût variable de l'enquête qui est directement relié au nombre d'unités échantillonnées. Le coût fixe, tel que le coût de la préparation du questionnaire et d'autres coûts au niveau organisationnel, ne fait pas partie de C_0 .

3.2 Plan à échantillons indépendants

Pour le plan à échantillons indépendants, la contrainte budgétaire est donnée par

$$C_0 = c_1^{\text{I}} n_1 + c_{\text{II}}^{\text{I}} n_1 m_1 + c_1^{\text{II}} n_2 + c_2^{\text{II}} n_2 m_2. \quad (3.1)$$

Les deux premiers termes sont les coûts de la première vague de collecte des données, et les deux derniers, ceux de la deuxième vague.

Proposition 4. Si les paramètres de configuration de l'enquête sont les mêmes aux deux périodes :

$$c_1^{\text{I}} = c_2^{\text{I}} = c_1^{\text{II}} = c_2^{\text{II}}, \quad c_{\text{II}}^{\text{I}} = c_{\text{II}}^{\text{II}} \quad (3.2)$$

alors les tailles optimales d'échantillon et les variances résultantes sont données par

$$m = \sqrt{\frac{c_1^{\text{I}}}{S_2^2} S_2^b - \frac{c_{\text{II}}^{\text{I}}}{S_2^2} / M}, \quad n = \frac{2 \{ c_1^{\text{I}} + [c_1^{\text{II}} S_2^b / (S_2^b - S_2^w / M)]^{1/2} \}}{C_0}, \quad V_{e,1}[d] = \frac{C_0}{4 \left[c_1^{\text{I}} + \sqrt{c_1^{\text{II}} S_2^w / (S_2^b - S_2^w / M)} \right]}$$

$$\times \left[S_2^b + \sqrt{\frac{c_1^{\text{I}}}{c_{\text{II}}^{\text{I}} S_2^b - S_2^w / M} - \frac{1}{S_2^w}} \right] - \frac{N}{2} S_2^b. \quad (3.3)$$

Dans les équations (3.3), les tailles d'échantillon n et m sont traitées comme des variables continues. En pratique, le nombre entier le plus proche doit être utilisé, une valeur d'au moins 2 étant nécessaire pour estimer la composante de

variance appropriée, et les valeurs maximales étant N et M , respectivement.

Le nombre d'unités d'observation échantillonnées dans une grappe dépend uniquement des coûts relatifs au niveau de la grappe et au niveau de l'unité d'observation, $c_1^{\text{I}}/c_{\text{II}}^{\text{I}}$, et des variances relatives S_2^b/S_2^w , ou, ce qui est équivalent, la corrélation intragrappe. Des coûts d'interview c_{II}^{I} plus élevés empêchent le concepteur de l'enquête d'utiliser un plus grand nombre d'unités d'observation : une augmentation de c_{II}^{I} entraîne une diminution de m ainsi que de n . Un coût au niveau de la grappe plus élevé donne lieu à une redistribution des unités échantillonnées : quand c_1^{I} augmente, n diminue tandis que m augmente. Une plus grande variance intragrappe S_2^w nécessite le tirage d'un plus grand nombre d'unités d'observation m dans une grappe pour maintenir la précision globale. Une plus grande variance intergrappes S_2^b nécessite l'échantillonnage d'un plus grand nombre n de grappes. Enfin, le budget total de l'enquête C_0 affecte le nombre de grappes n , mais non la taille m , du sous-échantillon. Par conséquent, la variance de d est inversement proportionnelle à C_0 .

Une situation non symétrique peut être traitée comme un sous-produit des conditions de premier ordre dérivées dans la preuve (voir l'annexe). Cependant, aucune solution analytique n'est disponible dans ce cas.

3.3 Plan à panel de grappes

Pour le plan à panel de grappes, la contrainte budgétaire est donnée par

$$C_0 = c_{12}^{\text{I}} n + c_{\text{II}}^{\text{I}} n m_1 + c_2^{\text{II}} n m_2. \quad (3.4)$$

Le premier terme est le coût au niveau de la grappe associé au plan d'échantillonnage, et les deux autres termes sont les coûts de la collecte des données au niveau individuel aux première et deuxième vagues, respectivement.

Proposition 5. Pour le plan à panel de grappes, les tailles d'échantillon sont données par

$$m_1 = 2C_0 / c_{\text{II}}^{\text{I}} \left(1 + \frac{\kappa S_2^{1w}}{S_2^b} + \sqrt{D} \right), \quad m_2 = \kappa m_1,$$

$$n = \frac{c_1^{\text{I}} + c_{\text{II}}^{\text{I}} m_1 + c_2^{\text{II}} m_2}{C_0}, \quad \kappa = \sqrt{\frac{c_{\text{II}}^{\text{I}} S_2^w}{c_1^{\text{I}} S_2^b}},$$

à condition que

(3.5)

Ici, l'indice inférieur c désigne le plan à panel de grappes (en anglais cluster) et p^i est la corrélation inter-temporelle, ou autocorrélation, des moyennes de grappe. L'indice supérieur i désigne le premier degré d'échantillonnage. Si p^i est positive, le plan à panel de grappes est plus efficace que le plan à échantillons indépendants pour des valeurs fixes de n et m . Sous les conditions de symétrie,

$$V_{e,c}[d] = 2 \left(1 - \frac{n}{N} \right) \frac{S_b^2(1 - p^i)}{S_b^2} + 2 \left(1 - \frac{m}{M} \right) \frac{nm}{S_b^2}, \quad (2.13)$$

où l'indice inférieur e , c signifie « variances égales, plan à panel de grappes ».

2.3 Plan à panel d'unités d'observation

Proposition 3. Soit n sur N grappes et m sur M unités d'observation échantillonnées sans remise à la première période et utilisées aux deux périodes. Alors

$$V_o[d] = \left(1 - \frac{n}{N} \right) \frac{S_{1b}^2 + S_{2b}^2 - 2\rho^i S_{1b} S_{2b}}{n} + \left(1 - \frac{m}{M} \right) \frac{S_{1w}^2 + S_{2w}^2 - 2\rho^{ii} S_{1w} S_{2w}}{nm}, \quad (2.14)$$

L'indice inférieur o désigne le « plan à panel d'unités d'observation ». Sous l'hypothèse des conditions symétriques,

$$V_{o,o}[d] = 2 \left(1 - \frac{n}{N} \right) \frac{n}{S_o^2(1 - p^i)} + 2 \left(1 - \frac{m}{M} \right) \frac{m}{(1 - p^{ii}) S_o^2},$$

$$p^{ii} = \frac{1}{N} \sum_{i=1}^N \sum_{j=1}^M (Y_{1ij} - \bar{Y}_{1i})(Y_{2ij} - \bar{Y}_{2i}), \quad (2.15)$$

où l'indice inférieur correspondant e , o signifie « variances égales, plan à panel d'unités d'observation ».

Ici, p^{ii} est la corrélation intertemporelle, ou autocorrélation, des observations individuelles dans les grappes. L'indice supérieur ii désigne le deuxième degré d'échantillonnage. Si p^{ii} est positive, le plan à panel d'unités d'observation est plus efficace que le plan à panel de grappes pour des valeurs fixes de n et m .

Comment sont reliées les deux autocorrélations qui figurent dans (2.15) ? Conceptuellement, nous pouvons imaginer n importe quel nombre de relations possibles entre elles. Introduisons un modèle de superpopulation

$$Y_{ij} = \mu_i + a_i + e_{ij}, \quad E_{\xi}[a_i] = 0, \quad E_{\xi}[e_{ij}] = 0, \quad (2.16)$$

3. Coûts pour des échantillons en grappes répétées

dans lequel les termes a_i et e_{sij} sont indépendants l'un de l'autre pour tout s , $t = 1, 2$. L'indice inférieur ξ désigne les espérances sous le modèle de superpopulation. Le cas où $p^i = 0$ et $p^{ii} = 1$ s'observe quand les variations des moyennes de grappes ont lieu indépendamment entre les grappes ($E_{\xi}[a_i a_{2i}] = 0$), mais que les individus gardent leur position à l'intérieur de la grappe, $e_{1ij} = e_{2ij}$. Le cas où $p^i = 1$ et $p^{ii} = 0$ a lieu quand les effets aléatoires de grappes sont les mêmes aux deux périodes, $a_{1i} = a_{2i}$, tandis que les effets aléatoires individuels ne sont pas corrélés ($E_{\xi}[e_{1ij} e_{2ij}] = 0$). Ni l'une ni l'autre de ces situations n'est entièrement réaliste. Cependant, on peut sans doute s'attendre à ce que la dynamique individuelle soit une source plus importante de variation au cours du temps que la dynamique de grappes, de sorte que les relations $p^{ii} \geq p^i \geq 0$ sont les plus plausibles. Nous examinerons dans les exemples numériques des sections 4 et 5 la mesure dans laquelle le choix du meilleur plan est sensible à la relation entre les deux corrélations.

À la présente section, nous analysons la rentabilité des échantillons en grappes lorsque l'on veut estimer la différence entre deux moyennes d'échantillon pour deux périodes différentes.

Une discussion des coûts de l'échantillonnage en grappes est donnée dans Kish (1995, section 8.3B), Thompson (1992, section 12.5), ainsi que Lehtonen et Pahkinen (2004). Un exposé mathématique plus détaillé, avec les formules de variance corrigées pour les populations finies, peut être consulté dans Hansen et coll. (1953, volume II, section 6.11).

3.1 Notation et modèles de coûts

Supposons la structure de coût suivante, qui est une extension de celle de Kish (1995) pour les enquêtes répétées :

- c_1^i est le coût au niveau de la grappe au temps $t = 1$ pour les grappes qui sont utilisées à la première vague seulement ;
- c_2^i est le coût au niveau de la grappe pour une nouvelle grappe au temps $t = 2$;
- c_1^{ii} est le coût au niveau de la grappe pour les grappes dans lesquelles les données sont recueillies aux deux périodes $t = 1$ et $t = 2$ (coût du panel d'UPB) ;
- c_2^{ii} est le coût au niveau individuel au temps $t = 1$ pour les individus qui sont observés à la première vague seulement ;
- c_2^{ii} est le coût au niveau individuel au temps $t = 2$ pour les individus qui sont observés à la deuxième vague seulement ;

y_{ij} dans l'échantillon. Les totaux de population $T[\cdot]$ et leurs estimations $t[\cdot]$ peuvent être calculés comme il suit :

total de grappe :

$$T''[Y] = \sum_{i=1}^I X_{i''}, t''[Y] = \frac{M}{N} \sum_{i=1}^I y_{i''},$$

total de population :

$$T''[Y] = \sum_{i=1}^I X_i, t''[Y] = \frac{n}{N} \sum_{i=1}^I t_i[Y]. \quad (2.1)$$

Les moyennes pour les unités d'observation sont

$$\bar{y}_{i''} = \frac{1}{M} \sum_{j=1}^M y_{ij} = \frac{T''_{ij}[Y]}{T''_{i''}[Y]}, \bar{y}_{i'} = \frac{1}{m} \sum_{j=1}^m y_{ij} = \frac{t''_{ij}[Y]}{t''_{i''}[Y]},$$

$$\bar{y}_{i''} = \frac{T''_{i''}[Y]}{\sum_{j=1}^M X_{ij}} = \frac{NM}{\sum_{j=1}^M \sum_{i=1}^n X_{ij}}, \bar{y}_{i'} = \frac{t''_{i''}[Y]}{\sum_{j=1}^m X_{ij}} = \frac{nm}{\sum_{j=1}^m \sum_{i=1}^n X_{ij}}. \quad (2.2)$$

La variance de X et ses composantes intragrappe (indiquée par l'indice w pour *within*) et intergrappe (indiquée par b pour *between-cluster*) sont

$$S^2_{i''} = \frac{\sum_{j=1}^M \sum_{i=1}^n (X_{ij} - \bar{X}_{i''})^2}{NM - 1}, \quad (2.3)$$

$$S^2_{i'w} = \frac{\sum_{j=1}^m (X_{ij} - \bar{X}_{i'})^2}{M - 1}, S^2_{i'w} = \frac{1}{N} \sum_{i=1}^n S^2_{i'w}, \quad (2.4)$$

$$S^2_{i''} = \frac{\sum_{j=1}^M (X_{ij} - \bar{X}_{i''})^2}{N - 1}. \quad (2.5)$$

La caractéristique d'intérêt principal est la variation des moyennes,

$$D = \bar{Y}_{i''} - \bar{Y}_{i'}, \quad (2.6)$$

estimée par

$$d = \bar{y}_{i''} - \bar{y}_{i'}. \quad (2.7)$$

Une propriété de cet estimateur intéressante pour les analystes et les utilisateurs des données est sa cohérence interne : l'estimateur de la différence est égal à la différence des estimateurs. Si les échantillons de périodes consécutives ne se chevauchent que partiellement, l'estimateur composite ou l'estimateur des MCG (Fuller 1999, Hansen, Hurwitz et Madow 1953, Patterson 1950, Rao et Graham 1964, Wolter 2007) sont plus efficaces.

Dans la suite de l'exposé, nous supposons que toutes les procédures d'échantillonnage correspondent à l'échantillonnage aléatoire simple sans remise. Pour la moyenne contemporeine, la variance est donnée par (Cochran 1977, Th. 10.1) :

$$V_c[d] = \left(1 - \frac{n}{N}\right) \frac{S^2_{i''} + S^2_{i'w} - 2\rho^1_{i''} S^2_{i''} S^2_{i'w}}{nm}, \quad \rho^1 = \frac{1}{N} \sum_{i=1}^I \frac{S^{1b}_{i''} S^{2b}_{i''} (N-1)}{\sum_{i=1}^I (Y_{i''} - \bar{Y}_{i''})(Y_{i''} - \bar{Y}_{i'})}. \quad (2.12)$$

Proposition 2. Soit n sur N grappes échantillonnées sans remise à la première période et utilisées aux deux périodes. Soit m sur M unités d'observation échantillonnées sans remise indépendamment aux deux périodes. Alors

2.2 Plan à panel de grappes

où l'indice inférieur e_i signifie « variances égales, plan à échantillons indépendants ».

$$V_{e_i}[d] = 2 \left(1 - \frac{n}{N}\right) \frac{S^2_{i''} + S^2_{i'w}}{nm} + 2 \left(1 - \frac{m}{M}\right) \frac{S^2_{i''}}{nm}, \quad (2.11)$$

Le résultat découle immédiatement de (2.8) en raison de l'indépendance des deux échantillons. L'indice inférieur i dants ». Sous les conditions de symétrie de (2.9), si les tailles d'échantillon sont les mêmes au deux périodes, $n_1 = n_2 = n$ et $m_1 = m_2 = m$, alors

$$V_i(d) = \left(1 - \frac{n_1}{N}\right) \frac{S^2_{i''}}{n_1} + \left(1 - \frac{n_2}{N}\right) \frac{S^2_{i''}}{n_2} + \left(1 - \frac{m_1}{M}\right) \frac{S^2_{i'w}}{m_1} + \left(1 - \frac{m_2}{M}\right) \frac{S^2_{i'w}}{m_2}. \quad (2.10)$$

Proposition 1. Soit le tirage sans remise de n_1 sur N grappes et de m_1 sur M unités d'observation dans les grappes au temps $t = 1$. Soit le tirage sans remise de n_2 sur N grappes et de m_2 sur M unités d'observation dans les grappes au temps $t = 2$, ce tirage étant indépendant de celui effectué au temps $t = 1$. Alors

2.1 Plan à échantillons indépendants

de travail raisonnable. Les dérivations analytiques sont possibles sans ces hypothèses, mais deviennent extrêmement lourdes. En outre, il n'est pas raisonnable de penser que le concepteur d'enquête pourrait connaître les caractéristiques de la future population. Donc (2.9) devrait être considérée comme un modèle

$$S^{1w}_{i''} = S^{2w}_{i''} = S^2_{i''}, S^{1w}_{i''} = S^{2w}_{i''} = S^2_{i''}, S^{1w}_{i''} = S^{2w}_{i''} = S^2_{i''}, S^{1w}_{i''} = S^{2w}_{i''} = S^2_{i''}. \quad (2.9)$$

Pour la simplicité et la clarté de l'exposé, nous formulerons souvent l'hypothèse de conditions symétriques :

$$V[\bar{y}_{i''}] = \left(1 - \frac{n}{N}\right) \frac{S^2_{i''}}{n} + \left(1 - \frac{m}{M}\right) \frac{S^2_{i'w}}{m}. \quad (2.8)$$

entre ces moyennes, pour les trois plans d'échantillonnage

étudiés.

Afin d'intégrer les aspects économiques de la collecte des données, à la section 3, nous présentons un modèle de coût relativement simple pour une enquête avec échantillonnage en grappes répétées. Nous définissons et résolvons les problèmes d'optimisation pour obtenir les tailles d'échantillon optimales pour les trois plans pris en considération. En introduisant dans le modèle les estimations des paramètres statistiques (variances et autocorrélations) et des composantes de coût (coûts au niveau de la grappe et au niveau individuel), le concepteur d'enquête peut comparer les valeurs numériques des variances pour choisir le meilleur plan. À la section 4, nous illustrons cette approche et montrons que chacun des plans pris en considération peut être le meilleur, selon les valeurs des paramètres. Les résultats intuitifs (par exemple, le coût plus élevé de la collecte des données et les autocorrélations plus faibles des caractéristiques observées rendent les modes de collecte des données par panel moins intéressants) sont assortis d'une justification analytique et d'un support quantitatif.

Les sections 2 à 4 traitent de l'efficacité de l'estimation de la différence entre les moyennes seulement, mais des objectifs de collecte des données plus réalistes comprennent l'estimation de caractéristiques contemporaines et de leurs variances. À cette fin, à la section 5, nous présentons un cadre de maximisation de l'utilité qui décrit le choix de scénarios d'échantillonnage qui s'offre au concepteur de l'enquête. Ce cadre fournit une fonction objectif agrégée qui combine plusieurs critères d'élaboration du plan. Les résultats sont de nouveau ceux attendus : si les modes de collecte des données par panel plus coûteux produisent des tailles d'échantillon plus petites, les estimations des moyennes sont moins efficaces que dans le cas de plans d'échantillonnage plus simples. Le seul moyen de justifier cette perte d'efficacité est d'obtenir une amélioration radicale de l'estimation de la différence, qui ne peut se produire que si les autocorrélations sont plus fortes. Les effets de ce genre sont également illustrés à la section 5. À la section 7, nous présentons nos conclusions. Les preuves sont données en annexe.

2. Variances sous le plan

Posons que la population est constituée de N grappes, ou UPE, aux deux périodes et que chaque grappe est constituée de M individus, ou USE. De cette population est tiré un EAS de $1 < n_i \leq N$ grappes au temps $t = 1, 2$, et un EAS de $1 < m_i \leq M$ individus dans chaque grappe présente dans l'échantillon au temps t . Soit l'indice i désignant les UPE et l'indice j désignant les USE. Donc, la mesure typique sera désignée X_{ij} dans la population et

chevauchement des échantillons, et tirer deux échantillons indépendants à deux périodes distinctes. Nous donnons à ce plan le nom de plan à *échantillons indépendants*. Il pourrait aussi juger utile de recycler les UPE d'une vague à l'autre. S'il estime qu'il est difficile de suivre les USE d'une vague à l'autre, les sous-échantillons dans les grappes peuvent être tirés indépendamment aux deux vagues de collecte des données. Nous donnerons à ce type de plan le nom de plan à *panel de grappes*. Si une très grande précision est essentielle, le plan entièrement longitudinal visera à repérer tous les individus qui ont répondu à la première vague et à leur demander de participer à une deuxième interview. Afin de distinguer ce plan de celui à panel de grappes, nous le nommerons plan à *panel d'unités d'observation*.

Un aspect particulier qui nous semble est important dans la gestion des enquêtes, mais qui n'est pas souvent abordé dans la littérature existante, est le coût de mise en œuvre (Groves 1989). Les modèles de coût classiques, tel celui utilisé pour établir le plan de répartition optimale de l'échantillon de Neyman-Tchuprow (Neyman 1938), peuvent être étendus afin d'y inclure des termes de coût pour la première visite à la grappe et à l'unité d'observation finale, ainsi que pour les visites subséquentes. En ce qui concerne les grappes, le coût est vraisemblablement plus faible à la deuxième visite. Il n'est plus nécessaire de créer de nouvelles cartes ni d'établir des listes. Les mêmes intervieweurs peuvent être choisis pour effectuer les interviews lors des vagues subséquentes de collecte des données. La coopération avec les dirigeants communautaires, parfois importante, par exemple chez certaines sociétés traditionnelles, a déjà été établie. L'effet de la collecte des données par panel au niveau individuel est moins évident. Un ménage interviewé lors des vagues précédentes qui a déménagé doit éventuellement être dépisté dans une région géographique différente, ce qui accroît le coût (moyen) d'interview du panel. La probabilité que des circonstances de ce genre se présentent augmente dans le cas typique d'intervalles plus longs entre les vagues de l'enquête dans les pays en développement : les intervalles entre les vagues de la DHS sont habituellement de cinq à sept ans. En revanche, si un mode d'interview moins coûteux peut être utilisé après la première vague (par exemple, le coût de l'interview téléphonique au lieu d'une visite sur place), le coût de l'interview du panel diminue.

Dans le présent article, nous tenons compte de considérations statistiques ainsi qu'économiques en vue de choisir le plan de sondage approprié et ses paramètres. Nous supposons que le concepteur de l'enquête souhaite estimer la variation de la moyenne de population entre deux périodes, et/ou les moyennes proprement dites. À la section 2, nous présentons une population sommaire et calculons les variances sous le plan des moyennes, ainsi que et la différence

Rentabilité des enquêtes avec échantillonnage en grappes répétées

Stanislav Kolenikov et Gustavo Angeles¹

Résumé

Nous analysons l'efficacité statistique et économique de diverses enquêtes avec échantillonnage en grappes pour lesquelles la collecte des données est effectuée à deux périodes, ou vagues, consécutives. Dans le cas d'un plan à échantillons indépendants, un échantillon en grappes est tiré de manière indépendante à chacune des deux vagues. Dans le cas d'un plan à panel de grappes, les mêmes grappes sont utilisées aux deux vagues, mais le tirage des échantillons dans les grappes est effectué indépendamment aux deux périodes. Dans un plan à panel d'unités d'observation, les grappes ainsi que les unités d'observation sont retenues d'une vague de collecte des données à l'autre. En supposant que la structure de la population est simple, nous calculons les variances sous le plan ainsi que les coûts des enquêtes réalisées selon ces divers types de plan. Nous considérons d'abord l'estimation de la variation de la moyenne de population entre deux périodes et nous déterminons les répartitions d'échantillon optimales pour les trois plans étudiés. Nous proposons ensuite un cadre de maximisation de l'utilité emprunté à la microéconomie en vue d'illustrer une approche possible pour choisir le plan dans laquelle nous nous efforçons d'optimiser simultanément plusieurs variances. La prise en compte simultanée de plusieurs moyennes et de leurs variances a tendance à faire pencher la préférence du plan à panel d'unités d'observation vers les plans à panel de grappes et échantillons indépendants plus simples si le mode de collecte de données par panel est trop coûteux. Nous présentons des exemples numériques qui illustrent comment un concepteur d'enquête pourrait choisir le plan efficace sachant les paramètres de population et les coûts de collecte des données.

Mots clés : Étude longitudinale ; échantillon en grappes ; DHS ; NHIS.

1. Introduction

Afin d'analyser la dynamique des phénomènes sociaux, comportementaux ou de santé des populations, les chercheurs et les responsables de l'élaboration des politiques doivent obtenir des renseignements sur les caractéristiques de la population à de multiples périodes. Les enquêtes à plan de sondage complexe sont la source d'information utilisée le plus fréquemment pour les grandes populations, telles que celle d'un pays dans son ensemble. Outre les aspects habituellement pris en considération dans les enquêtes ponctuelles, comme la stratification et la mise en grappes, d'autres éléments peuvent être importants dans les enquêtes dont les données sont recueillies pour deux périodes ou plus. Dans ce genre d'enquête, le coût total et l'erreur d'enquête totale sont influencés par le chevauchement entre les échantillons consécutifs, l'érosion (informative) de l'échantillon, les effets de la durée de la présence dans l'échantillon ou du conditionnement, et d'autres facteurs dynamiques.

En vue d'estimer les variations au moyen d'enquêtes répétées, il est souvent souhaitable que la corrélation temporelle entre les unités observées soient élevée, ce qui peut être réalisé en administrant le questionnaire de l'enquête aux mêmes unités d'échantillonnage et/ou d'observation. Dans les enquêtes longitudinales, une visite est faite aux mêmes unités d'observation (personnes, ménages) pendant plusieurs périodes, éventuellement pendant un nombre indéfini de périodes (Panel Study of Income Dynamics (PSID) des

Etats-Unis, British Household Panel Study (BHPS) et d'autres). Un recueil de renseignements sur les études longitudinales peut être consulté sur le site Web de l'Institute for Social and Economics Research, à l'adresse <http://iseressex.ac.uk/isc/keeptrack/index.php>. Dans les enquêtes à panel rotatif, les unités d'observation sont recrutées dans l'échantillon pour quelques périodes, puis sortent de l'échantillon et sont de nouveau étudiées à une période ultérieure. La Current Population Survey (CPS) des Etats-Unis (Binder et Hidiroglou 1988, Eckler 1955, Rao et Graham 1964) et un certain nombre d'enquêtes environnementales (Fuller 1999, McDonald 2003, Scott 1998) sont des exemples d'enquête à panel rotatif. Une autre option consiste à utiliser les mêmes unités primaires d'échantillonnage (UPE) aux différentes vagues, mais à sélectionner indépendamment les unités d'observation (unités secondaires d'échantillonnage, USE). Les enquêtes dont les données sont recueillies de cette façon comprennent les Demographic and Health Surveys (DHS) internationales et la National Health Interview Survey (NHIS) des Etats-Unis.

Nous nous concentrerons sur les enquêtes dont les données sont recueillies à deux périodes, ou vagues, en utilisant un plan d'échantillonnage en grappes à deux degrés à chaque vague de collecte des données. Nous considérons trois plans de sondage possibles, qui se distinguent par l'importance et la profondeur du chevauchement des unités d'échantillonnage au cours du temps. Le concepteur de l'enquête peut se limiter à réduire tout effet éventuel du

1. Stanislav Kolenikov, Department of Statistics, 146 Middlebush Hall, University of Missouri, Columbia, MO 65211-6100, U.S.A. Courriel : kolenikovs@missouri.edu; Gustavo Angeles, Associate Director of the Center for Evaluation Research, National Institute of Public Health, Mexico, Mexico. Courriel : gangelles@insp.mx.

En insérant (B.2) et (B.4) dans (B.1), nous avons

$$E_{\pi}(I_{KN}^{*}) = I_{KN}^{*} + \frac{1}{N^2} \sum_L \sum_{h=1}^H c_k^* \sum_{i \in A_h} w_i^* E_{\pi}(q_k^2) \pi_{i2}^{-2} (y_i - \bar{y}_{h2})^2 + o_p(n^{-1}),$$

et, comme $E_{\pi}(q_k^2) = p_k(1 - p_k)$, nous avons (20).

Bibliographie

- En insérant (B.2) et (B.4) dans (B.1), nous avons
- Berger, Y.G. (2007). A jackknife variance estimator for unistage stratified samples with unequal probabilities. *Biometrika*, 94, 953-964.
- Binder, D.A., Babyak, C., Brodeur, M., Hidiroglou, M. et Jocelyn, W. (2000). Variance estimation for two-phase stratified sampling. *The Canadian Journal of Statistics*, 28, 751-764.
- Breidt, F.J. et Fuller, W.A. (1993). Regression weighting for multipurpose samplings. *Sankhyā*, B, 55, 297-309.
- Brick, J.M., et Morganstein, D. (1996). WesVarPC: Software for computing variance estimates from complex designs. *Proceedings of the 1996 Annual Research Conference*, U.S. Bureau of the Census, 861-866.
- Cochran, W.G. (1977). *Sampling Techniques*. New York : John Wiley & Sons, Inc.
- Fay, R.E. (1984). Some properties of estimates of variance based on replication methods. *Proceedings of the Survey Research Section*, American Statistical Association, 495-500.
- Flyer, P. (1987). Finite population correction for replication estimates of variance. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 732-736.
- Fuller, W.A. (1998). Replication variance estimation for two-phase samples. *Statistica Sinica*, 8, 1153-1164.
- Fuller, W.A. (2003). Estimation for multiple phase samples. Dans *Analysis of Survey Data*, (Eds., R.L. Chambers et C.J. Skinner). Wiley, Chichester, England, 307-322.
- Wolter, K. (2007). *Introduction to Variance Estimation*. 2^{ème} Edition, New York : Springer.
- Hidiroglou, M.A. (2001). L'échantillonnage double. *Techniques d'enquête*, 27, 157-169.
- Hidiroglou, M.A., et Sæmdal, C.-E. (1998). Emploi des données auxiliaires dans l'échantillonnage à deux phases. *Techniques d'enquête*, 24, 11-20.
- Kim, J.K., Navarro, A. et Fuller, W.A. (2006). Replicate variance estimation after multi-phase stratified sampling. *Journal of the American Statistical Association*, 101, 312-320.
- Kim, J.K., et Sitter, R.R. (2003). Efficient variance estimation for two-phase sampling. *Statistica Sinica*, 13, 641-653.
- Kott, P.S., et Stukel, D.M. (1997). La méthode du jackknife convient-elle à un échantillon à deux degrés ? *Techniques d'enquête*, 23, 89-98.
- Lu, W., et Sitter, R.R. (2006). Risque de divulgation et estimation de la variance. *Recueil de la série des symposiums internationaux de Statistique Canada*, 11-522-XIF.
- Neyman, J. (1938). Contribution to the theory of sampling human populations. *Journal of the American Statistical Association*, 33, 101-116.
- Rao, J.N.K. (1965). On two simple schemes of unequal probability sampling without replacement. *Journal of the Indian Statistical Association*, 3, 173-180.
- Rao, J.N.K. (1973). On double sampling for stratification and analytical surveys. *Biometrika*, 60, 125-133.
- Rao, J.N.K., et Shao, J. (1992). Jackknife variance estimation with survey data under hot deck imputation. *Biometrika*, 79, 811-822.
- Rao, J.N.K., et Sitter, R.R. (1995). Variance estimation under two-phase sampling with application to imputation for missing data. *Biometrika*, 82, 453-460.
- Rust, K.F., et Rao, J.N.K. (1996). Variance estimation for complex surveys using replication techniques. *Statistical Methods in Medical Research*, 5, 283-310.
- Samford, M.R. (1967). On sampling without replacement with unequal probability of selection. *Biometrika*, 54, 499-513.

$$\text{Var} \left(\sum_{i \in U_h} \pi_{i2}^{-1} a_{i1} \right) = N_h^2 \sum_{i \in U_h} \pi_{i1}^{-2} (1 - \pi_{i2}) (Y_i - \bar{Y}_h)^2 + o(N^{-1}). \quad (\text{A.1})$$

Donc, le biais de l'estimateur de variance KNF est de la forme (13) sous l'hypothèse d'échantillonnage de Poisson de a_i .

B. Preuve de (20)

Pour chaque k ,

$$\bar{y}_{ip}^{*(k)} - \bar{y}_{ip} = \bar{y}_{ip}^{*(k)} - \bar{y}_{ip} + \bar{y}_{ip} - \bar{y}_{ip},$$

où $\bar{y}_{ip}^{*(k)}$ est défini par (12). Donc,

$$\begin{aligned} I_{*}^{\text{KNF}} &= \sum_{k=1}^K c_k (\bar{y}_{ip}^{*(k)} - \bar{y}_{ip})^2 = \sum_{k=1}^K c_k (\bar{y}_{ip}^{*(k)} - \bar{y}_{ip})^2 \\ &\quad + 2 \sum_{k=1}^K c_k (\bar{y}_{ip}^{*(k)} - \bar{y}_{ip}) (\bar{y}_{ip}^{*(k)} - \bar{y}_{ip}) \\ &\quad + \sum_{k=1}^K c_k (\bar{y}_{ip}^{*(k)} - \bar{y}_{ip})^2. \end{aligned} \quad (\text{B.1})$$

En vertu de la construction de $\bar{y}_{ip}^{*(k)}$, nous avons

$$E^*(\bar{y}_{ip}^{*(k)}) = \bar{y}_{ip}^{(k)} + o_p(n^{-1}). \quad (\text{B.2})$$

En outre, en écrivant $q_{ki} = M_{(k)}^{(2)} - 1$, nous avons $q_{ki} = O_p(n^{-1/2})$ et nous pouvons appliquer un développement en série de Taylor pour obtenir

$$\bar{y}_{ip}^{*(k)} = \bar{y}_{ip}^{(k)} + \frac{\sum_{i \in A_{h2}} w_i \pi_{i2}^{-1} q_{ki} (Y_i - \bar{y}_{h2})}{\sum_{i \in A_{h2}} w_i \pi_{i2}^{-1}} + o_p(n^{-1}). \quad (\text{B.3})$$

En outre, comme

$$\frac{1}{I} \sum_{h \in A_{h2}} w_i \pi_{i2}^{-1} z_i - \frac{1}{I} \sum_{h \in A_{h2}} w_i \pi_{i2}^{-1} z_i = O_p(n^{-1})$$

pour toute variable z dont les quatrième moments sont bornés, nous pouvons montrer que (B.3) se réduit à

$$\bar{y}_{h2}^{(k)} = \bar{y}_{h2}^{(k)} + \frac{\sum_{i \in A_{h2}} w_i \pi_{i2}^{-1} q_{ki} (Y_i - \bar{y}_{h2})}{\sum_{i \in A_{h2}} w_i \pi_{i2}^{-1}} + o_p(n^{-1}).$$

D'où, nous pouvons écrire

$$\sum_{k=1}^K c_k (\bar{y}_{ip}^{*(k)} - \bar{y}_{ip}^{(k)})^2 = \sum_{k=1}^K c_k (\bar{y}_{ip}^{(k)} - \bar{y}_{ip}^{(k)})^2$$

$$\left\{ \sum_{L=1}^K c_k \left(N^{-1} \sum_{h=1}^H \sum_{i \in A_{h2}} w_i \pi_{i2}^{-1} q_{ki} (Y_i - \bar{y}_{h2}) \right)^2 + o_p(n^{-1}) \right\}. \quad (\text{B.4})$$

Remerciements

Puisque la méthode de répliques proposée fournit des estimateurs de variance convergents pour les moyennes de population, elle peut être appliquée facilement à d'autres paramètres de population finie qui sont des fonctions lisses des moyennes de population.

Dans le cas de certaines grandes enquêtes, le nombre de répliques peut être assez élevé, parce que la méthode utilise le même nombre de répliques pour l'échantillon de première phase. Si l'on souhaite réduire le nombre de répliques, la méthode de Fuller (1998) ou celle de Kim et Sitter (2003) est une option. Un examen plus approfondi suivant cette piste sera le sujet d'une prochaine étude.

Annexe

A. Preuve de (13)

Soit $\mathbf{a} = (a_1, \dots, a_N)$ où a_i est la version étendue de l'indicateur d'échantillonnage de deuxième phase discutée dans Kim et coll. (2006). C'est-à-dire, $a_i = 1$, si l'unité i est sélectionnée dans l'échantillon de deuxième phase après avoir été sélectionnée dans l'échantillon de première phase et $a_i = 0$ autrement.

En vertu de l'hypothèse (9), conditionnellement à \mathbf{a} , nous avons

$$\sum_{k=1}^K c_k (\bar{y}_{h2}^{(k)} - \bar{y}_{h2})^2 = \text{Var}(\bar{y}_{h2} | \mathbf{a}) + o_p(n^{-1}).$$

Donc, le biais de $\sum_{k=1}^K c_k (\bar{y}_{h2}^{(k)} - \bar{y}_{h2})^2$ en tant qu'estimateur de $\text{Var}(\bar{y}_{h2})$ est égal, en ignorant les termes d'ordre $o_p(n^{-1})$, à

$$E\{\text{Var}(\bar{y}_{h2} | \mathbf{a})\} - \text{Var}(\bar{y}_{h2}) = \text{Var}\{E(\bar{y}_{h2} | \mathbf{a})\}.$$

En utilisant la définition étendue de a_i , nous avons

$$E(\bar{y}_{h2} | \mathbf{a}) = \frac{\sum_{i \in U_h} \pi_{i2}^{-1} a_{i1}}{\sum_{i \in U_h} \pi_{i2}^{-1} a_i}$$

et, en vertu de l'hypothèse d'échantillonnage de Poisson des a_i ,

Tableau 2
Biais relatif (BR) et coefficient de variation (CV) pour les estimateurs de variance (5 000 échantillons)

Méthode	Estimateur	Echantillon-nage de phase première	Echantillon-nage de phase deuxième	BR (%)	CV(%)
KNF	REE	EAS	EAS St.	-11,25	18,22
		Pois. St.	-9,56	18,67	
		RS St.	-7,75	15,35	
		EAS St.	-8,05	18,61	
	RS	Pois. St.	-9,03	20,84	
		RS St.	-5,73	17,27	
		EAS St.	-6,76	22,32	
		Pois. St.	-6,06	15,81	
	REG	EAS	EAS St.	-3,26	12,82
		RS St.	-3,26	12,82	
		RS	EAS St.	-4,17	21,74
		Pois. St.	-3,64	16,92	
Nouveau REE	EAS	EAS St.	0,09	18,23	
		Pois. St.	-1,23	19,70	
		RS St.	-0,04	16,06	
		EAS St.	0,78	19,78	
	RS	Pois. St.	-2,07	21,26	
		RS St.	1,00	17,67	
		EAS St.	-0,61	22,00	
		Pois. St.	-0,57	16,55	
	REG	EAS	EAS St.	-0,08	13,36
		RS St.	0,67	22,86	
		Pois. St.	-0,01	16,97	
		RS St.	0,59	14,02	

KNF :	Estimateur de variance de Kim et coll. (2006) sans répétitions supplémentaires pour la correction du biais
Nouveau :	Estimateur de variance proposé (16)
REE :	Estimateur à facteur d'extension pondéré (23)
REG :	Estimateur par la régression (27)
SRS :	Échantillonnage aléatoire simple
RS :	Échantillonnage de Rao-Sampford
EAS Str :	Échantillonnage aléatoire simple stratifié
Pois. Str. :	Échantillonnage de Poisson stratifié
RS Str. :	Échantillonnage de Rao-Sampford stratifié.

6. Conclusion

L'estimation de la variance par répliques sous échantillonnage à deux phases représente un problème pratique important dans les sondages et la méthode KNF est un outil utile à cet égard. Dans le présent article, nous proposons une méthode qui est une extension de la méthode KNF en ce sens qu'elle s'applique directement quand la fraction d'échantillonnage n'est pas négligeable, sans accroître le nombre de répliques. La méthode proposée s'applique aussi à l'échantillonnage de Poisson avec probabilités inégales dans chaque strate à l'étape de l'échantillonnage de deuxième phase. La théorie a été élaborée uniquement sous échantillonnage de Poisson à la deuxième phase, mais les résultats de simulation présentés à la section 5 indiquent que la méthode proposée donne des résultats raisonnablement bons pour d'autres plans d'échantillonnage avec probabilités inégales, tels que l'échantillonnage de Rao-Sampford.

Tableau I
Moyenne et variance des estimateurs ponctuels (5 000 échantillons)

Dans cette simulation, comme la fraction d'échantillon-nage de première phase n est pas négligeable ($n/N = 0.2$), l'estimateur de variance KNF sans répliques supplémentaires sous-estime la variance réelle et l'estimateur de variance proposé estime la variance avec un biais plus petit, inférieur à 3 % en valeur absolue dans tous les cas, ce qui est en harmonie avec la théorie des sections 3 et 4. La valeur absolue du biais relatif dans l'estimateur de variance KNF est grande car, même si la variance due à \mathbf{T}_{x_1} est estimée avec convergence dans (29), la variance due à β_2 est sous-estimée sans les répliques supplémentaires. Le biais relatif dans notre estimateur proposé est réduit parce que les répliques (18) créent une variation supplémentaire dans les poids de rééchantillonnage par l'ajout d'une perturbation δ_k tirée d'une loi choisie convenablement. Le CV de l'estimateur de variance proposé est un peu plus grand que celui de l'estimateur KNF, parce que le premier comporte un caractère aléatoire supplémentaire dû à la génération de δ_k d'après (15).

Le tableau 2 donne le biais relatif (BR) et le coefficient de variation (CV) des deux estimateurs de variance. Nous avons calculé les biais relatifs des estimateurs de variance en divisant le biais Monte Carlo de l'estimateur de variance par la variance Monte Carlo de l'estimateur ponctuel. Pour calculer le coefficient de variation de l'estimateur de variance, nous avons divisé l'erreur-type Monte Carlo de l'estimateur de variance par la moyenne Monte Carlo de l'échantillon fixe dans le premier cas.

résultats numériques du tableau 1. L'échantillonnage de Rao-Sampford à la deuxième phase est légèrement plus efficace que l'échantillonnage de Poisson en raison de la

Estimateur	Echantillon-	nage de première phase	Echantillon- nage de deuxième phase	Moyen- ne	Variance
REE	EAS	EAS Str.	10,0	0,0749	
		Pois. Str.	10,0	0,0784	
		RS Str.	10,0	0,0754	
	RS	EAS Str.	10,0	0,0768	
		Pois. Str.	10,0	0,0827	
		RS Str.	10,0	0,0781	
REG	EAS	EAS Str.	10,0	0,0540	
		Pois. Str.	10,0	0,0510	
		RS Str.	10,0	0,0495	
	RS	EAS Str.	10,0	0,0551	
		Pois. Str.	10,0	0,0531	
		RS Str.	10,0	0,0515	
Estimateur à facteur d'extension repondéré (23)					
REG :	Estimateur par la régression (27)				
EAS :	Echantillonnage aléatoire simple				
RS :	Echantillonnage de Rao-Sampford				
EAS Str. :	Echantillonnage aléatoire simple stratifié				
Pois. Str. :	Echantillon de Poisson stratifié				
RS Str. :	Echantillonnage de Rao-Sampford stratifié.				

aléatoire sous l'échantillonnage de Poisson, mais est fixe sous l'échantillonnage de Rao-Sampford.

Nous avons employé pour la simulation une structure factorielle $2 \times 3 \times 2$ avec trois facteurs, qui sont :

1. Echantillonnage pour l'échantillon de première phase contre échantillonnage de Rao-Sampford de taille $n=200$ en utilisant z_i comme mesure de taille.
2. Echantillonnage pour l'échantillon de deuxième phase (3) : Echantillonnage aléatoire stratifié de taille $n_h=25$, échantillonnage de Poisson stratifié avec taille d'échantillon prévue $n_h=25$ en utilisant q_i comme mesure de taille pour l'échantillonnage avec probabilités inégales, et échantillonnage de Rao-Sampford stratifié de taille $n_h=25$ en utilisant q_i comme mesure de taille pour l'échantillonnage avec probabilités inégales.

3. Méthodes d'estimation de la variance (2) : Estimateur KNF (11) sans répliques supplémentaires contre estimateur de variance proposé en utilisant (16), calculé en se basant sur la méthode du jackknife.

À partir de la population finie générée ci-dessus, nous avons produit $B = 5\,000$ échantillons Monte Carlo indépendants pour la simulation. Dans le cas des plans de sondage avec échantillonnage de Rao-Sampford dans la première phase, nous avons utilisé la méthode d'estimation jackknife de la variance proposée par Berger (2007), qui donne un estimateur convergent de la variance d'échantillonnage de première phase. Le paramètre d'intérêt est la moyenne de population de la variable y . Pour chaque échantillon Monte Carlo, nous avons calculé deux estimateurs ponctuels, à savoir le REE donné par (24) et l'estimateur par la régression (REG) donné par (28) en utilisant la variable auxiliaire $(1, x_i)$. Nous avons calculé les biais relatifs des estimateurs de variance en divisant le biais Monte Carlo de l'estimateur de variance par la variance Monte Carlo de l'estimateur ponctuel.

Le tableau 1 donne la moyenne et la variance des deux estimateurs ponctuels. Pour l'estimation ponctuelle, l'estimateur par la régression est significativement plus efficace que l'estimateur REE pour cette population, car la variable auxiliaire x est corrélée avec la variable étudiée y . La variance asymptotique théorique de l'estimateur par la régression sous échantillonnage aléatoire simple à la première phase et sous échantillonnage aléatoire stratifié à la deuxième phase est approximativement égale à

$$\left(\frac{1}{200} - \frac{1}{1\,000} \right) 8 + \left(\frac{1}{100} - \frac{1}{200} \right) 4 = 0,052$$

et la variance asymptotique théorique de l'estimateur REE pour le même plan d'échantillonnage est, approximativement, $(1/100 - 1/1\,000) 8 = 0,072$, qui concorde avec les

intégrer l'échantillonnage stratifié dans la deuxième phase si \mathbf{x}_i englobe le vecteur des indicateurs de strate.

En utilisant les arguments de la section 3, nous pouvons construire la k^e réplique pour $Y_{i,REG}^{(k)}$ par

$$Y_{i,REG}^{(k)} = \mathbf{T}_{x,1}^{(k)} \boldsymbol{\beta}_2^{(k)}, \quad (29)$$

où

$$\mathbf{T}_{x,1}^{(k)} = \sum_{i \in I_1} w_i^{(k)} \mathbf{x}_i$$

$$\boldsymbol{\beta}_2^{(k)} = \left(\sum_{i \in I_2} w_i^{(k)} w_{*2}^{(k)} \mathbf{x}_i' \mathbf{x}_i' \right)^{-1} \sum_{i \in I_2} w_i^{(k)} w_{*2}^{(k)} \mathbf{x}_i' y_i$$

et $w_{*2}^{(k)}$ est défini dans (27).

La méthode de répliques (29) peut être appliquée directement à l'estimateur par calage à deux phases dont ont discuté Hidiroglou et Särndal (1998). Si $H = 1$, alors la réplique de $\boldsymbol{\beta}_2$ donnée par (29) se réduit à

$$\hat{\boldsymbol{\beta}}_2^{(k)} = \left(\sum_{i \in I_2} w_i^{(k)} M_{i2}^{(k)} \pi_{i2}^{-1} \mathbf{x}_i' \mathbf{x}_i' \right)^{-1} \sum_{i \in I_2} w_i^{(k)} M_{i2}^{(k)} \pi_{i2}^{-1} \mathbf{x}_i' y_i$$

5. Étude par simulation

Àfin d'étudier la performance en échantillon fini des estimateurs proposés, nous avons procédé à une étude par simulation limitée. Dans la simulation, nous avons com-mencé par générer une population finie artificielle de taille $N = 1\,000$ avec cinq variables $(z_i, q_i, x_i, y_i, u_i)$, où les éléments de population sont générés indépendamment à partir de $z_i \sim \exp(1) + 2$; $q_i \sim \chi^2(1) + 2$; $x_i \sim N(2, 1)$; $u_i \sim \text{Unif}\{1, 2, 3, 4\}$, où $\text{Unif}\{1, \dots, G\}$ désigne une loi uniforme discrète avec support $\{1, \dots, G\}$; et

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 z_i + \beta_3 q_i + e_i$$

avec $(\beta_0, \beta_1, \beta_2, \beta_3) = (0, 2, 1, 1)$ et $e_i \sim N(0, 1)$. The variables z_i, q_i, x_i, u_i , et e_i sont mutuellement indépendantes. Nous avons défini la strate pour l'échantillon de deuxième phase en utilisant la variable u_i . Nous nous sommes servis de la variable x_i pour calculer l'estimateur par la régression à deux phases (28) avec $\mathbf{x}_i' = (1, x_i)'$, de la variable z_i comme mesure de taille pour l'échantillonnage avec probabilités inégales dans l'échantillonnage de première phase, et de la variable q_i comme mesure de taille pour l'échantillonnage avec probabilités inégales dans l'échantillonnage de deuxième phase.

Pour obtenir les échantillons sélectionnés avec probabilités inégales de cette étude par simulation, nous avons utilisé l'échantillonnage de Poisson ou l'échantillonnage de Rao-Sampford (Rao 1965 et Sampford 1967), avec des probabilités de sélection proportionnelles à la mesure de la variable de taille. Notons que la taille d'échantillon finale est

Sous les conditions de régularité discutées dans KNF,

nous avons

$$E(V_{KNF}^*) = V_{KNF}^* + N^{-2} \sum_{h=1}^H \sum_{i \in A_{h2}} w_i^2 b_i^2 \pi_{i2}^{-2} u(y_i - \bar{y}_{h2})^2 \quad (20)$$

$$+ o_p(n^{-1}), \quad (20)$$

où $u = \sum_{k=1}^{K-1} c_k p_k (1 - p_k)$. Une esquisse de preuve de (20) est présentée à l'annexe B. Si b_i est déterminé par

$$b_i = \sqrt{(1 - \pi_{i2}) w_i^{-1} u^{-1}}, \quad (21)$$

l'estimateur de variance (16) est convergent, parce que le deuxième terme de (20) annule V_{bias} dans (14). Il en est ainsi même quand la fraction d'échantillonnage de première phase n/N n'est pas négligeable. Pour garantir que les poids de rééchantillonnage soient non négatifs dans (18), nous exigeons que b_i dans (19) soit ≤ 1 . Si nous posons $p_k = 0.5$, alors

$$b_i = \sqrt{\frac{4(1 - \pi_{i2}) w_i^{-1}}{\sum_{k=1}^K c_k}},$$

qui est inférieur ou égal à 1 si $\sum_{k=1}^K c_k \geq 4$. En fait, les p_k peuvent être égales à n'importe quel nombre compris entre 0 et 1, à condition que le b_i résultant dans (21), soit inférieur ou égal à 1.

4. Extensions

À la présente section, nous considérons l'extension de la méthode de répliques proposée à d'autres types d'estimateurs à deux phases que le REE donné en (10).

4.1 Estimateur à facteur d'extension double

En échantillonnage à deux phases, on se sert également de l'estimateur à facteur d'extension double (*double expansion estimator*), qui doit son nom à Kot et Stukel (1997). L'estimateur à facteur d'extension double (DEB) possède la forme simple

$$\bar{y}_{DEB} = \frac{1}{H} \sum_{h=1}^H \sum_{i \in A_{h2}} w_i \pi_{i2}^{-1} y_i. \quad (22)$$

Quand l'échantillon de deuxième phase est un échantillon aléatoire stratifié, $\pi_{i2} = r_h/n_h$, et la méthode KNF peut être appliquée en utilisant la réplique

$$y_{DEB}^{(k)} = \frac{1}{H} \sum_{h=1}^H \left(\frac{\sum_{i \in A_{h1}} w_i^{(k)} w_i^{-1}}{\sum_{i \in A_{h2}} w_i^{(k)} w_i^{-1}} \right) \sum_{i \in A_{h2}} w_i^{(k)} y_i.$$

L'estimateur de variance KNF pour l'estimateur DEB est convergent quand la fraction d'échantillonnage de première

phase est négligeable. Si elle ne l'est pas, nous pouvons utiliser la méthode de répliques proposée à la section 3. La méthode de répliques proposée pour l'estimateur DEB crée les répliques

$$\bar{y}_{DEB}^{*(k)} = \frac{1}{H} \sum_{h=1}^H \sum_{i \in A_{h2}} w_i' w_i'^{-1} y_i', \quad (23)$$

où

$$w_{i2}^{*(k)} = M_{i2}^{(k)} \frac{\sum_{i \in A_{h1}} w_i^{(k)} w_i^{-1} M_{i2}^{(k)}}{\sum_{i \in A_{h1}} w_i^{(k)} w_i^{-1}},$$

et $M_{i2}^{(k)}$ est le facteur de répliques défini en (19). Le biais de l'estimateur de variance par répliques en utilisant la réplique (23) est négligeable si les répliques sont construites de façon à satisfaire (21).

Si l'échantillon de deuxième phase est obtenu par échantillonnage avec probabilités inégales dans chaque strate, la méthode de répliques telle que (23) ne peut pas être appliquée directement. L'estimateur DEB donné par (22) est généralement moins efficace que l'estimateur REE donné par (10). Notons que ce dernier peut être exprimé sous la forme

$$\bar{y}_{REE} = \frac{1}{H} \sum_{h=1}^H \sum_{i \in A_{h2}} w_i w_i^{-1} y_i, \quad (24)$$

où

$$w_i^* = \pi_{i2}^{-1} \frac{\sum_{i \in A_{h1}} w_i}{\sum_{i \in A_{h2}} w_i \pi_{i2}^{-1}}. \quad (25)$$

Les répliques (17) peuvent s'écrire

$$\bar{y}_{REE}^{*(k)} = \frac{1}{H} \sum_{h=1}^H \sum_{i \in A_{h2}} w_i^{(k)} w_{i2}^{*(k)} y_i', \quad (26)$$

où

$$w_{i2}^{*(k)} = M_{i2}^{(k)} \pi_{i2}^{-1} \frac{\sum_{i \in A_{h2}} w_i^{(k)} M_{i2}^{(k)} \pi_{i2}^{-1}}{\sum_{i \in A_{h1}} w_i^{(k)} M_{i2}^{(k)} \pi_{i2}^{-1}} \quad (27)$$

et $M_{i2}^{(k)}$ est défini dans (19).

4.2 Estimateur par la régression

En échantillonnage à deux phases, les variables auxiliaires observées dans l'échantillon de première phase peuvent être utilisées à l'étape de l'estimation. L'estimateur par la régression à deux phases du total de population peut s'écrire sous la forme

$$\bar{y}_{REG} = \bar{\mathbf{T}}_{x1}^T \hat{\boldsymbol{\beta}}_2 \quad (28)$$

où $\bar{\mathbf{T}}_{x1} = \sum_{i \in A_1} w_i \mathbf{x}_i$ est le vecteur des totaux de population de la variable de contrôle \mathbf{x}_i estimés sur l'échantillon de première phase, $\hat{\boldsymbol{\beta}}_2 = (\sum_{i \in A_2} w_i w_i^* \mathbf{x}_i \mathbf{x}_i^T)^{-1} \sum_{i \in A_2} w_i w_i^* \mathbf{x}_i y_i$ est un vecteur de coefficients de régression estimés sur l'échantillon de deuxième phase et w_i^* est donné par (25). Notons que l'estimateur par la régression donné par (28) peut

À la présente section, nous envisageons également un cas plus intéressant d'échantillonnage stratifié avec probabilités inégales pour la deuxième phase. Plus précisément, nous considérons que l'échantillon de deuxième phase est un échantillon de Poisson avec probabilités inégales dans les strates de deuxième phase. Fuller (1998) a également examiné l'échantillonnage de Poisson à la deuxième phase et soutenu qu'il s'agit d'une bonne approximation. Un exemple de cette situation dans le contexte des enquêtes sur les forêts est que, en plus des types de forêt, les interprètes des photos peuvent aussi déterminer, d'après les photos aériennes prises à la première phase, la densité des arbres et la hauteur des arbres, qui peuvent être utilisées pour construire les probabilités de sélection de deuxième phase dans chaque strate (type de forêt).

À la présente section, nous nous concentrons d'abord sur l'estimateur de type RFE, puisqu'il est plus efficace que celui de type DEF, et nous discutons de l'extension à l'estimateur DEF à la section 4. Soit w_i , le poids d'échantillonnage de première phase et soit w_{i2} , l'inverse la probabilité conditionnelle à la deuxième phase. C'est-à-dire, $w_{i2} = \pi_{i2}^{-1}$ où $\pi_{i2} = \Pr(i \in A_{h2} | i \in A_{h1})$. L'estimateur de type RFE peut s'écrire

$$\bar{y}_{ip}^* = \frac{1}{N} \sum_{h=1}^H N_{h1} \bar{y}_{h2} \quad (10)$$

où $N_{h1} = \sum_{i \in A_{h1}} w_i$ et $\bar{y}_{h2} = (\sum_{i \in A_{h2}} w_i \pi_{i2}^{-1})^{-1} \sum_{i \in A_{h2}} w_i \pi_{i2}^{-1} y_i$. Dans KNF, π_{i2} est supposée constante dans la strate de deuxième

phase. Nous considérons une approche fondée sur les répliques pour l'estimation de la variance de l'estimateur de type RFE (10) quand π_{i2} n'est pas nécessairement constante dans la strate de deuxième phase. Nous considérons le cas particulier où le plan d'échantillonnage de deuxième phase est un échantillonnage de Poisson. En utilisant la méthode de répliques satisfaisant (9), l'estimateur de variance de type KNF peut être appliqué à l'estimation de la variance de \bar{y}_{ip}^* dans cette situation. Autrement dit,

$$V_{KNF}^* = \sum_{k=1}^K c_k (\bar{y}_{ip}^{*(k)} - \bar{y}_{ip}^*)^2, \quad (11)$$

où

$$\bar{y}_{ip}^{*(k)} = \frac{1}{N} \sum_{h=1}^H N_{h1}^{(k)} \bar{y}_{h2}^{(k)} \quad (12)$$

avec $\bar{y}_{h2}^{(k)} = (\sum_{i \in A_{h2}} w_i^{(k)} \pi_{i2}^{-1})^{-1} \sum_{i \in A_{h2}} w_i^{(k)} \pi_{i2}^{-1} y_i$ et $N_{h1}^{(k)} = \sum_{i \in A_{h1}} w_i^{(k)}$, et c_k est un facteur associé à la réplique k déterminé par la méthode de répliques. Sous échantillonnage de Poisson à la deuxième phase, nous avons le biais asymptotique suivant :

$$\text{Biais}(V_{KNF}^*) = -\frac{1}{N^2} \sum_{h=1}^H \sum_{i \in U_h} \pi_{i2} (1 - \pi_{i2}) (y_i - \bar{y}_{h2})^2, \quad (13)$$

À la présente section, nous envisageons également un cas plus intéressant d'échantillonnage stratifié avec probabilités inégales pour la deuxième phase. Plus précisément, nous considérons que l'échantillon de deuxième phase est un échantillon de Poisson avec probabilités inégales dans les strates de deuxième phase. Fuller (1998) a également examiné l'échantillonnage de Poisson à la deuxième phase et soutenu qu'il s'agit d'une bonne approximation. Un exemple de cette situation dans le contexte des enquêtes sur les forêts est que, en plus des types de forêt, les interprètes des photos peuvent aussi déterminer, d'après les photos aériennes prises à la première phase, la densité des arbres et la hauteur des arbres, qui peuvent être utilisées pour construire les probabilités de sélection de deuxième phase dans chaque strate (type de forêt).

À la présente section, nous nous concentrons d'abord sur l'estimateur de type RFE, puisqu'il est plus efficace que celui de type DEF, et nous discutons de l'extension à l'estimateur DEF à la section 4. Soit w_i , le poids d'échantillonnage de première phase et soit w_{i2} , l'inverse la probabilité conditionnelle à la deuxième phase. C'est-à-dire, $w_{i2} = \pi_{i2}^{-1}$ où $\pi_{i2} = \Pr(i \in A_{h2} | i \in A_{h1})$. L'estimateur de type RFE peut s'écrire

$$V_{KNF}^* = \sum_{k=1}^K c_k (\bar{y}_{ip}^{*(k)} - \bar{y}_{ip}^*)^2 \quad (16)$$

où

$$\bar{y}_{ip}^{*(k)} = \frac{1}{N} \sum_{h=1}^H N_{h1}^{(k)} \bar{y}_{h2}^{(k)} \quad (17)$$

avec $N_{h1}^{(k)} = \sum_{i \in A_{h1}} w_i^{(k)}$, et $\bar{y}_{h2}^{(k)} = (\sum_{i \in A_{h2}} w_i^{(k)} \pi_{i2}^{-1})^{-1} \sum_{i \in A_{h2}} w_i^{(k)} \pi_{i2}^{-1} y_i$. Dans KNF, π_{i2} est supposée constante dans la strate de deuxième

$$M_{h2}^{(k)} = 1 + (\delta_{h2} - p_k) b_i \quad (19)$$

et b_i doit également être déterminé. Par construction, $E_*(\delta_{h2} - p_k) = 0$, où E_* indique que l'espérance est considérée par rapport au mécanisme précisé dans (15). Donc, les répliques (18) créent une variation supplémentaire dans les poids de rééchantillonnage, la variation supplémentaire (18) provenant de la distribution (15). Un choix approprié de p_i et b_i peut rendre convergent l'estimateur de variance résultant.

où $\bar{y}_2 = \sum_{h=1}^H w_h \bar{y}_{h2}$. L'estimateur de variance (3) est un estimateur de variance linéarisé. Kott et Stukel (1997) et KNF ont élaboré un estimateur de variance jackknife en supprimant successivement des unités de l'échantillon complet de première phase, puis en ajustant les poids. Les répliques jackknife complètes sont

$$\bar{y}_{dp}^{(k)} = \frac{1}{H} \sum_{h=1}^H N_{h1}^{(k)} \bar{y}_{h2}^{(k)} \quad (4)$$

où k est l'indice de l'unité supprimée dans la réplique jackknife.

$$\frac{1}{N} N_{h1}^{(k)} = \sum_{i \in A_{h1}} w_i^{(k)}$$

$$= \begin{cases} (n-1)^{-1} (n_h-1) & \text{si } k \in A_{h1} \\ (n-1)^{-1} n_h & \text{si } k \notin A_{h1} \end{cases}$$

et

$$\bar{y}_{h2}^{(k)} = \frac{\sum_{i \in A_{h2}} w_i^{(k)} y_i}{\sum_{i \in A_{h2}} w_i^{(k)}}$$

$$= \begin{cases} (r_h-1)^{-1} (r_h \bar{y}_{h2} - y^k) & \text{si } k \in A_{h2} \\ \bar{y}_{h2} & \text{si } k \notin A_{h2} \end{cases} \quad (5)$$

L'estimateur de variance jackknife complet de la forme

$$V_j = \sum_{k \in A_1} \frac{n}{n-1} (1 - f_1) (\bar{y}_{dp}^{(k)} - \bar{y}_{dp})^2, \quad (6)$$

où $\bar{y}_{dp}^{(k)}$ est défini dans (4), est asymptotiquement équivalent à

$$V_j \doteq n^{-1} (1 - f_1) \sum_H w_h (\bar{y}_{h2} - \bar{y}_2)^2 + (1 - f_1) \sum_H r_h^{-1} w_h^2 s_{h2}^2. \quad (7)$$

Donc, si nous comparons (7) à (2), le biais de l'estimateur de variance jackknife (6) est

$$\text{Biais}(V_j) \doteq -E \left\{ f_1 \sum_H (r_h^{-1} - n_h^{-1}) s_{h2}^2 \right\}.$$

Par conséquent, si la fraction d'échantillonnage de première phase est négligeable au sens où $f_1 \doteq 0$, le biais est négligeable, c'est-à-dire que le biais $= o(n^{-1})$. Sinon, l'estimateur de variance sous-estime la variance. Afin d'examiner une méthode du jackknife corrigée du biais, au lieu de (5), nous considérons

$$\bar{y}_{h2}^{(k)} = \begin{cases} (r_h - \delta_h)^{-1} (r_h \bar{y}_{h2} - \delta_h y^k) & \text{si } k \in A_{h2} \\ \bar{y}_{h2} & \text{si } k \notin A_{h2} \end{cases} \quad (8)$$

Ici L est le nombre de répliques. Pour la plupart des plans de sondage mesurables, c'est-à-dire les plans dont toutes les probabilités d'inclusion conjointe sont positives, nous pouvons construire l'estimateur de variance par répliques qui satisfait (9) même quand la fraction d'échantillonnage $f = n/N$ est grande. Par exemple, voir Fay (1984) et Flyer (1987). Brick et Morganstein (1996) décrivent l'algorithme de base pour WesVar, un logiciel disponible dans le commerce pour l'estimation de la variance par répliques dans le cas des sondages.

$$\frac{\text{Var}(\hat{\theta})}{V_1} - 1 = o_p(1). \quad (9)$$

seul (de premier) degré. Autrement dit, pour la variance de $\hat{\theta}$ sous le plan d'échantillonnage à un où $\hat{\theta} = \sum_{i \in A_1} w_i y_i$, et $\hat{\theta}^{(k)} = \sum_{i \in A_1} w_i^{(k)} y_i$, est convergent

$$V_1 = \sum_{k=1}^L c_k (\hat{\theta}^{(k)} - \hat{\theta})^2,$$

que l'estimateur de variance par répliques de la forme plus général. Pour cela, nous devons émettre l'hypothèse Nous étendons maintenant la méthode proposée à la

3. Méthode proposée

l'hypothèse que $f_1 \doteq 0$. résultant est approximativement sans biais sans émettre né de cette façon dans (8), l'estimateur de variance jackknife où $d_h = \sqrt{(1 - f_1 r_h n_h^{-1}) / (1 - f_1)}$. D'où, avec δ_h déterminé

$$\delta_h = \frac{1 + \sqrt{r_h (r_h - 1) / d_h}}{r_h}$$

Le biais asymptotique est nul si

$$\text{Biais}(V_j) \doteq E \left[\sum_{h=1}^H \left\{ (1 - f_1) \frac{(r_h - 1) \delta_h^2}{r_h^2} - \frac{1}{r_h} \left(1 - f_1 \frac{r_h n_h}{r_h} \right) \right\} w_h^2 s_{h2}^2 \right].$$

Donc, le biais asymptotique est donné par

$$+ (1 - f_1) \sum_H (r_h - 1) \delta_h^2 w_h^2 s_{h2}^2.$$

$$V_j \doteq n^{-1} (1 - f_1) \sum_H w_h (\bar{y}_{h2} - \bar{y}_2)^2$$

(5) est asymptotiquement équivalent à

L'estimateur de variance jackknife en utilisant (8) au lieu de où δ_h doit être déterminé. In (5), nous avons utilisé $\delta_h = 1$.

h . À la deuxième phase, un échantillon aléatoire stratifié de taille r est sélectionné en tirant un échantillon de taille $r_h (\leq n_h)$ dans la strate h , où $r = \sum_{h=1}^H r_h$ et la fraction d'échantillonnage r_h/n_h est fixe pour chaque strate. Pour discuter formellement de la théorie asymptotique, nous supposons que nous avons une suite de populations finies, une suite d'échantillons de première phase et une suite d'échantillons de deuxième phase, comme il est décrit dans KNF. Dans ces conditions asymptotiques, nous permettons que la taille de l'échantillon de deuxième phase r tende vers l'infini à la même vitesse que la taille de l'échantillon de première phase n , c'est-à-dire $r = O(n)$ et $r^{-1} = O(n^{-1})$, et H est fixe. Donc, dans les conditions où H est fixe, $r_h^{-1} = O(n^{-1})$.

Quand la variable étudiée y_i est observée dans l'échantillon de deuxième phase, la moyenne de population de y est estimée par

$$\bar{y}_{ip} = \frac{1}{H} \sum_{h=1}^H \sum_{i \in A_{h2}} \frac{n_h}{r_h} y_i,$$

où A_{h2} est l'ensemble d'indices pour les éléments de l'échantillon de deuxième phase qui appartiennent à la strate h . La variance de \bar{y}_{ip} peut s'écrire

$$\text{Var}(\bar{y}_{ip}) = \left(\frac{1}{H} - \frac{1}{N} \right) S_2^2 + E \left[\sum_{h=1}^H \left(\frac{n_h}{n} \right)^2 \left(\frac{1}{1} - \frac{r_h}{n_h} \right) S_{h1}^2 \right] \quad (1)$$

où $\bar{y}_1 = n^{-1} \sum_{h=1}^H \sum_{i \in A_{h1}} y_i$, $S_2^2 = (N - 1)^{-1} \sum_{i=1}^N (y_i - \bar{y})^2$, $S_{h1}^2 = (n_h - 1)^{-1} \sum_{i \in A_{h1}} (y_i - \bar{y}_{h1})^2$, et $\bar{y}_{h1} = n_h^{-1} \sum_{i \in A_{h1}} y_i$. En utilisant

$$n^{-1} S_2^2 = E \left\{ n^{-1} \sum_{h=1}^H w_h [(\bar{y}_{h1} - \bar{y})^2 + s_{h1}^2] \right\}$$

où $w_h = n^{-1} n_h$ et \bar{y} indique une approximation en ignorant les termes d'ordre $o(n^{-1})$, le terme de variance (1) est approximé par

$$\text{Var}(\bar{y}_{ip}) = E \left\{ n^{-1} (1 - \frac{1}{N}) \sum_{h=1}^H w_h (f_1 - \bar{y}_1)^2 \right. \\ \left. + \sum_{h=1}^H (r_h^{-1} - n_h^{-1} f_1) w_h s_{h1}^2 \right\}, \quad (2)$$

où $f_1 = nN^{-1}$.

Un estimateur convergent de la variance de \bar{y}_{ip} peut être dérivé de (2) en remplaçant \bar{y}_{h1} et s_{h1}^2 par leurs estimations respectivement. Autrement dit, un estimateur de variance convergent est donné par

$$\hat{V} = n^{-1} (1 - \frac{1}{N}) \sum_{h=1}^H w_h (\bar{y}_{h2} - \bar{y})^2 \\ + \sum_{h=1}^H (r_h^{-1} - n_h^{-1} f_1) w_h s_{h2}^2, \quad (3)$$

2. Conditions de base

Le plan de l'article est le suivant. À la section 2, nous présentons les conditions de base et à la section 3, nous décrivons la méthode proposée. À la section 4, nous étendons la méthode proposée à d'autres estimateurs sous échantillonnage à deux phases. À la section 5, nous présentons les résultats d'une étude par simulation limitée. Enfin, à la section 6, nous présentons nos conclusions.

Dans le présent article, nous proposons une nouvelle méthode de répliques pour l'estimation de la variance sous échantillonnage à deux phases. La méthode proposée est une extension de la méthode KNF qui englobe la situation où la fraction d'échantillonnage de première phase n est pas nécessairement négligeable. Contrairement à la méthode KNF, la méthode proposée ne requiert pas de répliques supplémentaires pour la correction du biais dans l'estimation de la variance, mais nécessite des corrections des poids de rééchantillonnage. En outre, la méthode proposée est applicable à l'échantillonnage de Poisson avec probabilités inégales dans les strates de deuxième phase, lequel n'a pas été discuté dans KNF. La méthode fondée sur des répliques est très facile à mettre en œuvre et peut être appliquée à divers types d'estimateurs.

(2006, KNF) offrent une étude rigoureuse de la méthode de répliques et l'envisagent pour d'autres types d'estimateurs. La méthode KNF a été élaborée principalement dans les conditions où la fraction d'échantillonnage de première phase est négligeable et l'échantillonnage de deuxième phase est un échantillon aléatoire stratifié. Si la fraction d'échantillonnage de première phase n est pas négligeable, des répliques supplémentaires sont nécessaires pour obtenir des estimations de variance convergentes.

Pour mieux exposer la motivation de l'étude, à la présente section, nous supposons simplement que l'échantillonnage de première phase correspond au tirage d'un échantillon aléatoire simple de taille n d'une population finie de taille N et que l'échantillon de deuxième phase est un échantillon aléatoire stratifié. À la section 3, nous étendons les conditions afin d'inclure tout échantillonnage arbitrairement mesurable à la première phase et l'échantillonnage de Poisson avec probabilités inégales dans chaque strate à la deuxième phase. Au moyen de l'information obtenue auprès de l'échantillon de première phase, ce dernier est stratifié en H strates pour l'échantillonnage de deuxième phase. Dans la strate h , nous avons n_h éléments de l'échantillon de première phase et soit A_{h1} l'ensemble d'indices pour les éléments de l'échantillon de première phase dans la strate

Estimation de la variance par répliques sous échantillonnage à deux phases

Jae Kwang Kim et Cindy Long Yu¹

Résumé

Dans l'échantillonnage à deux phases pour la stratification, l'échantillon de deuxième phase est sélectionné selon un plan stratifié basé sur l'information observée sur l'échantillon de première phase. Nous élaborons un estimateur de variance corrigé du biais fondé sur une méthode de répliques qui étend la méthode de Kim, Navarro et Fuller (2006). La méthode proposée est également applicable quand la fraction d'échantillonnage de première phase n'est pas négligeable et que le tirage de l'échantillon de deuxième phase se fait par échantillonnage de Poisson avec probabilités inégales dans chaque strate. La méthode proposée peut être étendue à l'estimation de la variance pour les estimateurs par la régression à deux phases. Les résultats d'une étude par simulation limitée sont présentés.

Mots clés : Échantillonnage double ; jackknife ; estimateur à facteur d'extension
répondère.

1. Introduction

L'échantillonnage à deux phases, introduit pour la première fois par Neyman (1938) et parfois appelé échantillonnage double, est une méthode d'échantillonnage rentable. On y recourt habituellement quand la collecte des données sur les variables d'intérêt coûte très cher, mais qu'il est relativement peu coûteux de recueillir des données sur des variables corrélées aux variables d'intérêt. L'échantillonnage à deux phases a des applications sous diverses formes (par exemple, Rao 1973 ; Cochran 1977 ; Breidt et Fuller 1993 ; Rao et Sitter 1995 ; Hidiroglou et Samdal 1998 ; Fuller 1998 ; Hidiroglou 2001 ; Fuller 2003). L'échantillonnage à deux phases pour la stratification fait référence à la situation où l'observation provenant de l'échantillon de première phase est utilisée pour procéder à une stratification pour l'échantillonnage de deuxième phase. Si l'on sélectionne l'échantillon de première phase pour les besoins de la stratification, l'échantillonnage de deuxième phase représente un outil utile lorsqu'il n'existe pas au départ de base de sondage pour la stratification. Par exemple, dans le cas des enquêtes sur les forêts, il est très difficile et coûteux de se rendre dans les régions éloignées pour faire des déterminations sur place. Par contre, les photographies aériennes sont relativement peu coûteuses, et les déterminations ayant trait, par exemple, au type de forêt d'après les photos aériennes sont fortement corrélées à celles faites au sol et peuvent être utilisées pour stratifier l'échantillon de première phase.

L'estimation de la variance par répliques est très répandue dans le cas des enquêtes complexes. Rust et Rao (1996) et Wolter (2007) donnent deux aperçus complets de ce sujet. La méthode de répliques ne requiert pas le calcul de la

dérivée partielle du développement en série de Taylor et l'utilisateur peut facilement produire des estimations de variance sans connaître le plan d'échantillonnage qui a été utilisé pour recueillir les données. De surcroît, la tendance à recourir à cette méthode s'accroît à cause des questions de confidentialité (Lu et Sitter 2006). Une fois que les poids de rééchantillonnage sont fournis, il n'est pas nécessaire que l'utilisateur dispose d'informations sur le plan de sondage, telles que les identificateurs de strate, pour analyser les données.

Deux estimateurs de la moyenne de population sous échantillonnage à deux phases sont utilisés fréquemment, à savoir l'estimateur à facteur d'extension double (DEF pour double expansion estimator) et l'estimateur à facteur d'extension répondère (RBE pour reweighted expansion estimator) qui doivent leur nom à Kott et Stukel (1997). En général le RBE est plus efficace que le DEF dans la situation d'échantillonnage à deux phases pour la stratification quand les y à l'intérieur d'une strate sont homogènes. L'estimation de la variance pour l'échantillonnage à deux phases est un problème compliqué en pratique et les praticiens s'intéressent aux méthodes de répliques pour l'estimation de la variance. Rao et Shao (1992) ont proposé un estimateur de variance par le jackknife convergent pour le RBE dans le contexte de l'imputation hot deck en traitant les répondants comme un échantillon de deuxième phase. Kott et Stukel (1997) ont examiné le même problème et concluent que les estimateurs de variance jackknife donnent de bons résultats pour le RBE si la fraction d'échantillonnage de première phase est négligeable. La fraction d'échantillonnage, $f_1 = nV^{-1}$, est dite négligeable si f_1 converge vers zéro sous les conditions asymptotiques décrites à la section 2. Binder, Babyak, Brodeur, Hidiroglou

- Kozak, M. (2004). Optimal stratification using random search method in agricultural surveys. *Statistics in Transition*, 6, 797-806.
- Kozak, M., et Verna, M.R. (2006). Approche de la stratification par une méthode géométrique et par optimisation : une comparaison de l'efficacité. *Techniques d'enquête*, 32, 177-183.
- Lavallée, P., et Hidiroglou, M. (1988). Sur la stratification de populations asymétriques. *Techniques d'enquête*, 14, 35-45.
- McEvoy, R.H. (1956). Variation in bank asset portfolios. *The Journal of Finance*, 11(4), 463-473.
- Rivest, L.-P. (1999). Stratum jumpers: Can we avoid them?. *ASA Proceedings of the Section on Survey Research Methods*, American Statistical Association, (Alexandria, VA), 64-72.
- Rivest, L.-P. (2002). Une généralisation de l'algorithme de Lavallée et Hidiroglou pour la stratification dans les enquêtes auprès des entreprises. *Techniques d'enquête*, 28, 207-214.
- Sæmdal, C.E., Swensson, B. et Wretman, J. (1992). *Model Assisted Survey Sampling*. New York : Springer Verlag.
- Sethi, V.K. (1963). A note on the optimum stratification of populations for estimating the population means. *Australian Journal of Statistics*, 5, 20-33.
- Sigman, R.S., et Monsour, N.J. (1995). Selecting samples from list frames of businesses. Dans *Business Survey Methods*, (Eds., B.G. Cox, D.A. Binder, B.N. Chinnappa, A. Christianson, M.L. Colledge et P.S. Kott), 133-152.
- Slamta, J., et Krenzke, T. (1996). Utilisation de la méthode de Lavallée et Hidiroglou pour le calcul des limites de stratification aux fins de l'enquête annuelle sur les dépenses en capital du Bureau of the Census. *Techniques d'enquête*, 22, 65-75.
- Sweet, E.M., et Sigman R.S. (1995). Evaluation of model-assisted procedures for stratifying skewed populations using auxiliary data. *U.S. Bureau of the Census* (www.census.gov/srd/papers/pdf/sm95-22.pdf).

Tableau 13
Tableau récapitulatif du programme R stratification

Argument	Strata.cumrootf	Strata.geo	Strata.IH	Strata.bh	Var.strata	Description	Format	Valeur par défaut
x	•	•	•	•		variable de stratification	vecteur	aucune (x est obligatoire)
n	•	•	•	•		taille totale d'échantillon cible	scalaire	aucune (n ou CV est obligatoire)
CV	•	•	•	•		CV ou EQ/MR cible	scalaire	aucune (n ou CV est obligatoire)
Ls	•	•	•	•		nombre de strates échantillonnées	scalaire	3
alloc	•	•	•	•		régle d'allocation (1)	liste (q1, q2, q3) ou q1 > 0	Neyman (q1=q3=0.5, q2=0)
certain	•	•	•	•		indices des x pour les unités à échantillonner obligatoirement	vecteur	NUTL (pas de strate à tirage obligatoire)
nclass	•					nombre de classes	scalaire	min(10L, N)
bh		•	•			bornes des strates	vecteur	aucune (bh est obligatoire)
takall.adjust			•			indicateur de correction pour strates à tirage complet	vecteur	FALSE (aucune correction)
takall	•	•	•			nombre de strates à tirage complet	un nombre parmi {0, 1, ..., Ls - 1}	0
initbh	•	•	•			bornes initiales des strates (2)	vecteur	bornes équidistantes
algo		•	•			identification de l'algorithme	liste (maxIter, method, minNh, "Kozak" ou "Sethi"	"Kozak"
algo.control			•			spécification des paramètres de l'algorithme (3)	maxstep, maxctl1, rep)	maxstep, maxctl1, rep)
strata	•	•	•	•		plan stratifié	objet de classe strata	aucune (strata est obligatoire)
y	•	•	•	•		variable d'étude	vecteur	NUTL (un modèle doit être fourni)
model	•	•	•	•		identification du modèle	"none", "loglinear", "none"	"none"
model.control	•	•	•	•		spécification des paramètres du modèle (4)	liste (beta, sig2, ph, ptkenone, gamma, epsilon)	depend du modèle, mais équivaut à model="none"
rh	•	•	•	•		taux de réponse anticipés	scalaire ou vecteur	rep (1, Ls) ou rh tiré de strata
rh.postcorr				•		indicateur de correction postérieure pour non-réponse	TRUE ou FALSE	FALSE (aucune correction)
takenone	•	•	•	•		nombre de strates à tirage nul	0 ou 1	0
bias.penalty	•	•	•	•		penalité pour le biais	scalaire	1

Bibliographie

Anderson, D.W., Kish, L. et Cornell, R.G. (1976). Quantifying gains from stratification for optimum and approximately optimum strata using a bivariate normal model. *Journal of the American Statistical Association*, 71, 887-892.

Baillargeon, S., Rivest, L.-P. et Ferland, M. (2007). Stratification en enquêtes entreprises : une revue et quelques avancées. *Recueil de la Section des méthodes d'enquêtes, Société Statistique du Canada* (www.ssc.ca/survey/documents/SSSC2007_S_Baillargeon.pdf).

Baillargeon, S., et Rivest, L.-P. (2009). A general algorithm for univariate stratification. *Revue Internationale de Statistique*, 77, 331-344.

Cochran, W.G. (1961). Comparison of methods for determining stratum boundaries. *Bulletin of the International Statistical Institute*, 32, 345-358.

Cochran, W.G. (1977). *Sampling Techniques*. Troisième Édition. New York : John Wiley & Sons, Inc.

Dalenius, T., et Hodges, J.L., Jr. (1959). Minimum variance stratification. *Journal of the American Statistical Association*, 54, 88-101.

Dayal, S. (1985). Allocation of sample using values of auxiliary characteristics. *Journal of Statistical Planning and Inference*, 11, 321-328.

DeLise, R.E., et Veum, C.S. (1991). Design issues for the Retail Trade Sample Surveys of the U.S. Bureau of the Census. *Proceedings of the Survey Research Methods Section, American Statistical Association*, 214-219.

Gunning, P., et Horgan, J.M. (2004). Un nouvel algorithme pour la construction de bornes de stratification dans les populations asymétriques. *Techniques d'enquête*, 30, 177-185.

Gunning, P., et Horgan, J.M. (2007). Improving the Lavallée and Hidroglou algorithm for stratification of skewed populations. *Journal of Statistical Computation and Simulation*, 77, 277-291.

Hidroglou, M.A. (1986). The construction of a self-representing stratum of large units in survey design. *The American Statistician*, 40, 27-31.

Hidroglou, M.A., et Sinha, K.P. (1993). Problems associated with designing subannual business surveys. *Journal of Business and Economic Statistics*, 11, 397-405.

Khan, M.G.M., Nand, N. et Ahmad, N. (2008). Détermination des bornes optimales de strate au moyen de la programmation dynamique. *Techniques d'enquête*, 34, 227-236.

(3) Les éléments à spécifier dans l'argument algo.control dépendent de l'algorithme. Le tableau qui suit indique les éléments employés par chaque algorithme avec leurs valeurs par défaut. On trouvera dans help(strata.LH) une description complète de chaque élément.

Algorithme	maxiter	method	minlh	maxstep	maxstll	rep	minsol
Sethi	500	-	-	-	-	3	1 000
Kozak initial	10 000	"original"	2	3	100	3	1 000
Kozak modifié	3 000	"modified"	2	3	-	-	1 000

(4) Les éléments de l'argument model.control dépendent du modèle :

- modèle loglinéaire avec mortalité (Loglinear) :

$$Y = \begin{cases} 0 & \text{avec probabilité } 1-p_h \\ \exp(\alpha + \beta \log(X) + \epsilon_{\text{psilon}}) & \text{avec probabilité } p_h \end{cases}$$

où $\epsilon_{\text{psilon}} \sim N(0, \text{sig2})$ est indépendant de X . Le paramètre p_h est spécifié par ph, plakenone et pcertain.

- modèle de remplacement aléatoire (random) :
$$Y = \begin{cases} X & \text{avec probabilité } 1 - \epsilon_{\text{psilon}} \\ X_{\text{new}} & \text{avec probabilité } \epsilon_{\text{psilon}} \end{cases}$$

où X_{new} est une variable aléatoire indépendante de X avec la même distribution que X .
- modèle linéaire hétéroscédastique (linear) :

$$Y = \beta X + \epsilon_{\text{psilon}}$$

où

$$\epsilon_{\text{psilon}} \sim N(0, \text{sig2} X^{\gamma_{\text{gamma}}}).$$

Le tableau qui suit présente les valeurs par défaut de model.control selon le modèle.

Modèle	beta	sig2	ph	plakenone	pcertain	gamma	epsilon
"Loglinear"	1	0	rep(1,1s)	1	-	-	-
"linear"	1	0	-	-	-	0	-
"random"	-	-	-	-	-	-	0

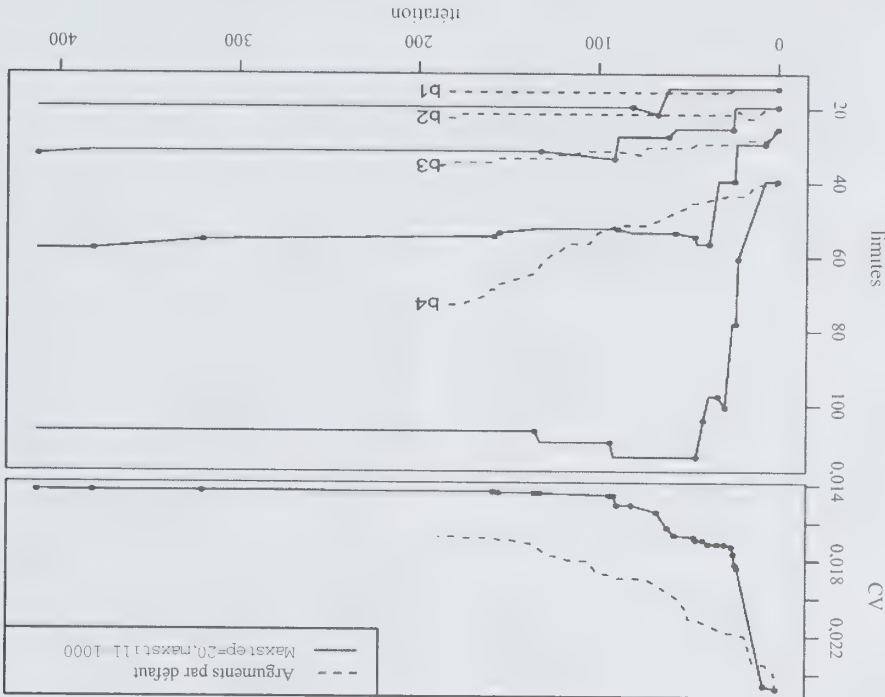


Figure 3 Succession des itérations pour deux exécutions de l'algorithme de Kozak

L'EQR de \bar{y}_s , mais aussi faire en sorte que les strates à tirage partiel contiennent au moins m_{\min} unités et que les tailles d'échantillon soient positives. La valeur par défaut de m_{\min} est 2. Nous disposons également d'un algorithme de Kozak à traitement non aléatoire avec `method="modified"` dans l'argument `algo.control`. Il essaie toutes les modifications de bornes possibles à une itération et sélectionne celle qui fait le plus baisser le critère d'optimisation. Il est plus lent que l'algorithme de Kozak sans améliorer pour autant la détection du minimum global de ce critère. Nous ne discuterons donc pas davantage de cette approche.

Pour illustrer ce qu'est le traitement complet des solutions possibles mentionné à la section 3.3, considérons l'ensemble de données `USbanks` qui contient 357 valeurs, dont seulement 200 sont uniques. Si nous voulons répartir cette population en deux strates, il n'y a que $\binom{200-1}{2-1} = 199$ solutions qui soient possibles. Avec la commande qui suit, nous pouvons énumérer toutes les solutions possibles :

```
> enum <- strata.LH(x = USbanks, CV = 0.05, Ls = 2,
  alloc = c(0.5, 0, 0.5))
```

On peut trouver ces solutions avec leur valeur correspondante de critère d'optimisation dans `enumsol.detail`. Seules les solutions satisfaisant aux contraintes d'admissibilité précitées y sont incluses.

Dans l'exécution de l'algorithme de Kozak, les valeurs initiales des bornes pourraient ne pas respecter les contraintes d'admissibilité et l'algorithme risquerait alors de ne pas fonctionner du tout. Dans ce cas, les bornes initiales seraient remplacées par des bornes robustes. Celles-ci donnent une strate à tirage nul vide si une telle strate est demandée. Les strates à tirage complet sont les plus petites possible et les strates à tirage partiel ont à peu près le même nombre de valeurs uniques de X .

Prenons une fois de plus l'exemple de la section 3.2 avec l'ensemble de données `UScities` où l'algorithme de Kozak a trouvé un minimum local avec les arguments par défaut. S'il utilise les bornes initiales géométriques, cet algorithme converge rapidement vers ce qui semble être un minimum global.

```
> LH_init <- strata.LH(x = UScities, initbh = pop2sbh,
  n = 100, Ls = 5, alloc = c(0.5, 0, 0.5), takeall = 0,
  algo.control = list(rep = 1))
> LH_init$iter.detail
```

	b1	b2	b3	b4	opt	step	iter	run
1	18.5	33.5	59.5	107	0.0144981	0	0	1
2	20.5	33.5	59.5	107	0.0143576	2	2	1
3	19.5	33.5	59.5	107	0.01434272	-1	10	1
4	19.5	33.5	58.0	107	0.01432714	-1	12	1
5	19.5	31.5	58.0	107	0.01431013	-2	13	1
6	19.5	32.5	58.0	107	0.01430163	1	63	1

L'élément `LH_init$iter.detail` donne en sortie les bornes initiales et celles des cinq itérations avec modification de bornes seulement. En tout 163 itérations ont été

```
> LH_param <- strata.LH(x = UScities, n = 100, Ls = 5,
  alloc = c(0.5, 0, 0.5), takeall = 0, algo.control =
  list(maxstep = 20, maxstill = 1000, rep = 20))
```

Les résultats des 20 répétitions figurent dans `LH_param$rep.detail` et sont résumés au tableau 12. La solution obtenue avec les bornes initiales géométriques se réalise 9 fois sur 20.

Tableau 12
Solutions obtenues par l'algorithme de Kozak pour 20 répétitions

CV	B1	B2	B3	B4	Fréquence
0.0143	19.50	32.50	58.00	107.00	9
0.0167	16.50	23.50	37.50	78.00	5
0.0167	15.50	22.50	35.50	73.00	6

La figure 3 montre que, par un accroissement des pas, l'algorithme peut plus facilement atteindre le minimum global ($CV=0,0143$) que s'il emploie les arguments par défaut (pointillés, $CV=0,0167$).

7.2 Tableau récapitulatif du programme R stratification

Dans cette annexe, nous résumons le programme *stratification*. Le tableau 13 énumère les cinq fonctions constitutives avec leurs arguments. Les notes qui suivent complètent le tableau.

(1) Dans la formule générale d'allocation (Hidiroglou et Srinath 1993), les tailles d'échantillon des strates sont proportionnelles à $N_{2q}^h \bar{Y}_{2q}^h S_{2q}^{yh}$.

(2) La valeur par défaut de `initbh` est l'ensemble de points de départ arithmétiques de Gunning et Horgan (2007) ; voir la section 3.3 à ce sujet. Avec `takenone=1` et si `initbh` est de la taille `Ls-1`, la borne initiale de la strate à tirage nul est fixée au premier percentile de X . Si ce dernier est égal à la valeur minimale de X , cette borne initiale donnerait une strate à tirage nul vide. Dans ce cas on fixe plutôt la borne initiale de la strate à tirage nul à la deuxième plus petite valeur de X .

obtient une valeur de n moindre que dans le traitement *a posteriori*. Au tableau 3 de Baillargeon et Rivest (2009), on trouvera d'autres exemples, incluant les moments antécipés ainsi que la non-réponse, de l'élaboration de plans stratifiés pour la population MRTS.

Tableau 10
Deux exemples de correction de non-réponse soit *a posteriori* (post) soit dans l'élaboration du plan

Méthode	rh	1	2	3	4	n	CV	anticipé
---------	----	---	---	---	---	-----	----	----------

cum \sqrt{f}	aucun	N_h	778	742	355	125	88	125
		n_h	87	90	98	125	390	1.11
		N_h^{post}	109	113	98	125	445	1.00
		N_h	778	742	355	125		
LH	donné	N_h	105	108	106	125	444	1.00
		n_h	105	108	106	125		
		N_h	774	675	374	177	379	1.11
		n_h^{post}	96	81	67	177	421	1.00
	donné	N_h	675	677	449	199	418	1.00
		n_h	70	69	80	199		
		N_h	675	677	449	199		
		n_h	70	69	80	199		

Une strate à tirage nul qui n'est pas échantillonnée peut présenter un avantage lorsqu'une population comporte de petites unités dont les valeurs Y sont proches de 0. On mesure alors la précision de \bar{y}_s par l'erreur quadratique moyenne $\text{Var}(\bar{y}_s) + (T_{0y}/N)^2$, où T_{0y} est le total prévu pour Y dans la strate à tirage nul. En prenant $\text{takenone}=1$ dans la fonction `strata.LH`, on se trouve à élaborer un plan optimal avec une strate à tirage nul. Baillargeon et Rivest (2009) ont démontré que l'algorithme de Sethi ne fonctionne pas en pareil cas et qu'il faut utiliser l'algorithme de Kozak. Si on fait intervenir une strate à tirage nul, on peut apporter une correction brute de biais en divisant \bar{y}_s par la proportion du total de la variable X dans les strates à tirage partiel. Ainsi, la pénalisation de biais dans l'erreur quadratique moyenne pourrait être trop sévère et une autre mesure de précision comme $\text{Var}(\bar{y}_s) + (P \times T_{0y}/N)^2$, pourrait être utilisée dans l'algorithme de stratification, P étant un chiffre dans l'intervalle (0, 1). Cette pénalisation de biais plus légère peut s'opérer en posant `bias.penalty` égal à P . Le code R qui suit élabore trois plans stratifiés optimaux pour la population MRTS avec et sans strate à tirage nul : la pleine pénalisation pour le biais par défaut est alors à comparer à une pénalisation plus légère avec $P = 0,5$.

```
> data(MRTS)
> notn <- strata.LH(x = MRTS, CV = 0.1, Ls = 3,
+ alloc = c(0.5, 0, 0.5))
> n1 <- strata.LH(x = MRTS, CV = 0.1, Ls = 3,
+ alloc = c(0.5, 0, 0.5), takenone = 1)
> n0.5 <- strata.LH(x = MRTS, CV = 0.1, Ls = 3,
+ alloc = c(0.5, 0, 0.5), takenone = 1, bias.penalty = 0.5)
```

5.2 Strate à tirage nul

On pourra voir au tableau 11 les tailles d'échantillon n pour les trois plans. Si on prévoit une strate à tirage nul avec pleine pénalisation de biais, n baisse de 22 à 16. Dans ce plan, la strate à tirage nul rend compte de 3 % du total pour la variable X . En ramenant la pénalisation pour le biais à $P = 0,5$, on augmente la taille de la strate à tirage nul et diminue n . D'autres exemples cités au tableau 2 de Baillargeon et Rivest (2009) montrent que la taille de la strate à tirage nul décroît habituellement avec l'EQMR cible. Dans l'exemple MRTS, l'ajout d'une telle strate réduit largement la valeur de n , alors que le plan ne change pas dans d'autres cas.

Tableau 11
Tailles d'échantillon de trois plans stratifiés optimaux pour la population MRTS

takenone=	bias.penalty=	n	% T_x	Conclusion		
				0	3	9
0	S.O.	22	16	13		
0	1	1				

Le programme R *stratification* offre des méthodes souples permettant d'élaborer des plans d'échantillonnage stratifiés à l'aide d'une variable unidimensionnelle de stratification comme une mesure de taille dans une enquête auprès des entreprises. Nous disposons de plusieurs méthodes pour établir les bornes et les tailles d'échantillon des strates. Ce programme permet de considérer plusieurs caractéristiques comme les strates à tirage complet ou à tirage nul, le degré de non-correspondance entre X et Y et une non-réponse par strate.

Remerciements

Nous sommes reconnaissants envers S. Fr, E. Gagnon, M. Kozak et J. Stardom pour toutes leurs observations constructives concernant ce programme de stratification, ainsi qu'à la Chaire de recherche du Canada en échantillonnage statistique et en analyse de données et au Conseil de recherches en sciences naturelles et en génie du Canada de l'aide financière qu'ils ont apportée. Ces travaux ont été financés par une bourse du U.S. National Science Foundation (SES-0751671).

7. Annexe

7.1 Description plus détaillée de l'algorithme de Kozak

Comme nous l'avons indiqué à la section 3.3, l'algorithme de Kozak effectue une recherche aléatoire des bornes. Pour être acceptée, une nouvelle borne doit non seulement faire baisser le critère d'optimisation, soit n ou

5. Autres caractéristiques

Baillargeon et Rivest (2009) se sont attachés à d'autres

aspects d'un plan stratifié, soit aux taux de non-réponse prévus par strate et à l'ajout d'une strate à tirage nul qui n'est pas échantillonnée. Dans cette section, nous examinerons brièvement comment ces aspects sont pris en charge par *stratification*. On se doit de tenir compte de la non-réponse dans le calcul de n . Une strate à tirage nul fait que \bar{y}_s est entaché d'un biais, auquel cas on spécifie le degré de précision cible sous forme d'erreur quadratique moyenne relative (EQMR) plutôt que de CV. La formule (4.3) de Baillargeon et Rivest (2009) donne une généralisation de (1) comprenant ces deux caractéristiques. C'est cette formule qui est utilisée pour le calcul des tailles d'échantillon dans la procédure d'optimisation.

5.1 Non-réponse

On peut apporter une correction de non-réponse *a posteriori* en divisant les tailles d'échantillon de strates sans non-réponse par les taux de réponse. C'est ce qui est illustré par le code R qui suit où figure la variable MRTS représentative de l'Enquête mensuelle sur le commerce de détail de Statistique Canada. On apporte postérieurement les corrections de non-réponse dans la fonction `var.strata` avec l'argument `rh.postcor=TRUE`. Une autre possibilité est de tenir compte des taux de réponse dans la répartition de l'échantillon entre les strates, ce qui peut être spécifié dans une fonction `strata` par l'argument `rh=`. Avec cette manière de procéder, on pénalise les strates à taux élevé de non-réponse, puisqu'on se retrouve normalement avec une valeur de n moindre que dans le cas des corrections *a posteriori*. C'est ce qui est illustré dans la partie cum \sqrt{f} du tableau 10. Avec quatre strates et des taux de réponse respectifs de 0,8, 0,8, 0,9 et 1, la correction *a posteriori* doit viser $n = 445$ pour la réalisation du CV cible comparativement à $n = 444$ pour une attribution tenant compte de la non-réponse.

```
> data(MRTS)
> cum <- strata.cumrootf(x = MRTS, nclass = 500, CV = 0.01,
  ls = 4, ailoc = c(0.5, 0, 0.5))
> cum.var <- var.strata(cum, rh = c(0.8, 0.8, 0.9, 1))
> cum.post <- var.strata(cum, rh = c(0.8, 0.8, 0.9, 1),
  rh.postcor = TRUE)
> cum.rh <- strata.cumrootf(x = MRTS, nclass = 500, CV = 0.01,
  ls = 4, ailoc = c(0.5, 0, 0.5), rh = c(0.8, 0.8, 0.9, 1))
```

On peut également tenir compte de la non-réponse lorsqu'on élabore un plan stratifié optimal soit *a posteriori* soit dans la construction des strates. Ces deux stratégies sont présentées pour la population MRTS dans le code R qui suit. Les taux élevés de non-réponse pour les petites unités pénalisent la première strate qui sera moindre en cas de prise en compte de la non-réponse dans l'algorithme de stratification, comme on peut le voir au tableau 10. En tenant compte de la non-réponse dans la construction des strates, on

4.2 Exemple tiré de Anderson, Kish et Cornell (1976) avec la distribution normale à deux variables

```
[1] 0.0689368
> geo_rml2 <- var.strata(geo_cer, model = "loglinear",
  model.control = list(beta = 1.058355, sig2 = 0.25677^2))
```

Dans Anderson *et coll.* (1976), on examine une stratification optimale pour Y étant donné X lorsque (X, Y) suit une distribution normale à deux variables avec corrélation ρ . Ceci correspond au modèle (2) avec $\alpha = \gamma = 0$, $\beta = \rho$, et $\sigma^2 = 1 - \rho^2$, où X a une distribution $N(0, 1)$. Pour reproduire les résultats de Anderson *et coll.* (1976), nous tirons une population de taille $N = 10^5$ d'une distribution $N(0, 1)$ et sélectionnons `model="linear"` (comme à la section 3.1, une moyenne de 10 empêche X de prendre des valeurs négatives). Dans le cas d'un modèle linéaire, seul l'algorithme de Kozak fonctionne. Ce problème étant partiel, nous réglons à 20 le paramètre `maxstep` et n'exécutons qu'une répétition (`rep=1`) de l'algorithme. En cas d'absence de strate à tirage complet, les limites optimales des strates sont indépendantes du CV comme à la section 3.1. Nous avons choisi CV = 0,01 dans les calculs.

```
> x <- rnorm(1e+05, 10)
> b13a <- strata.LH(x = x, CV = 0.01, ls = 3, takenone = 0,
  model = "linear",
  model.control = list(beta = 0.25, sig2 = 1 - 0.25^2,
  gamma = 0), algo.control = list(maxstep = 20, rep = 1)
> b13asbh - 10
[1] -0.619354 0.604198
```

Au tableau 9, les résultats du programme *stratification* sont égaux à ceux de Anderson *et coll.* (1976) jusqu'à près de deux décimales. Cela fait bien voir la polyvalence de ce programme qui peut trouver le plan stratifié optimal pour toute distribution de la variable de stratification et pour un certain nombre de modèles généraux pour la distribution conditionnelle de Y étant donné X .

Tableau 9 Comparaison des bornes optimales de strates dans Anderson *et coll.* (1976) et des bornes approximatives obtenues avec *stratification*

L	p	Résultats de stratification				Anderson et coll.			
		1	2	3	4	1	2	3	4
3	0,250	-0,619	0,604			-0,61	0,61		
	0,950	-0,591	0,568			-0,58	0,58		
	0,990	-0,571	0,549			-0,56	0,56		
	0,250	-0,984	0,004	0,985		-0,98	0,00	0,98	
	0,950	-0,930	0,009	0,942		-0,93	0,00	0,93	
4	0,250	-0,900	-0,001	0,895		-0,90	0,00	0,90	
	0,950	-0,902	-0,001	0,895		-0,90	0,00	0,90	
	0,250	-1,245	-0,377	0,387	1,251	-1,24	-0,38	0,38	1,24
	0,950	-1,187	-0,358	0,372	1,197	-1,19	-0,37	0,37	1,19
	0,990	-1,136	-0,344	0,353	1,144	-1,14	-0,35	0,35	1,14

Le modèle influe seulement sur les CV prévus. Tel n'est pas le cas pour le plan optimal où les moments anticipés sont employés dans l'algorithme de stratification. L'algorithme de Kozak pourrait ne pas trouver le minimum global n dans ce cas. Nous utilisons donc les bornes calculées avec X comme valeurs initiales.

```
> geo.cer.m <- strata.geo(x = X[ord], CV = 0.05, ls = 4,
  list(beta = 1.058355, sig2 = 0.256772))
> geo.cer.var <- var.strata(geo.cer.m, Y = Y[ord])
> cum.cer.m <- strata.cumcof(x = X[ord], nclass = 50,
  certain = (length(X) - 2) : length(X), model = "loglinear",
  model.control = list(beta = 1.058355, sig2 = 0.256772))
> LH <- strata.LH(x = X, CV = 0.05, ls = 5,
  allloc = c(0.35, 0.35, 0), takeall = 1)
> LH.var <- var.strata(LH, Y = Y)
> LH.m <- strata.LH(x = X, CV = 0.05, ls = 5,
  initlbh = LH$bh, allloc = c(0.35, 0.35, 0), takeall = 1,
  model = "loglinear", model.control = list(beta = 1.058355,
  sig2 = 0.256772))
> LH.m.var <- var.strata(LH.m, Y = Y)
```

Au tableau 8, les tailles d'échantillon calculées avec les moments anticipés donnent des CV de moins de 5 % pour l'estimation de la moyenne de la variable RM_{T85} . Le plan optimal LH exige un n un peu inférieur à ceux des deux autres. Si on tient compte de $Y \neq X$ quand on minimise (1), on obtient une strate à tirage complet plus grande, sa taille passant de 4 à 5.

Tableau 8
Trois plans stratifiés pour l'estimation de la moyenne de RM_{T85} avec REV_{84} comme variable de stratification

Méthode	Modèle					CV	anticipé
	1	2	3	4	5		
LH	n_h	N_h	n_h	N_h	n_h	4,90	
	127	79	46	29	3		
géométrique	n_h	N_h	n_h	N_h	n_h	4,74	
	121	81	45	32	5		
loglinéaire cum \sqrt{f}	n_h	N_h	n_h	N_h	n_h	4,78	
	42	116	88	35	3		
LH	n_h	N_h	n_h	N_h	n_h	7,11	
	127	79	46	29	3		
géométrique	n_h	N_h	n_h	N_h	n_h	6,89	
	120	82	45	33	4		
cum \sqrt{f}	n_h	N_h	n_h	N_h	n_h	7,37	
	127	79	46	29	3		

Notons enfin que les arguments `model` et `model.control` peuvent être employés avec `var.strata`. Pour le plan géométrique considéré ici, on peut obtenir des résultats fort semblables à ceux qu'on obtient avec l'argument $y=x$. Tel que montré ci-dessous, le modèle donne un CV de 6,894 % comparativement à 6,890 % pour la variable originale RM_{T85} . Dans le cas de la méthode cum \sqrt{f} , le CV du modèle est de 7,282 % par rapport à 7,369 % calculé plus haut. Pour l'algorithme de Lavallée et Hidiroglou, les deux valeurs s'établissent à 7,080 % et 7,110 %.

```
> data(Sweden)
> X <- Sweden$REV84
> Y <- Sweden$RM85
> ord <- order(X)
> geo.mnt <- var.strata(geo.cer, Y = Y[ord])
> cum.mnt <- var.strata(cum.cer, Y = Y[ord])
> c(geo.mnt$RMSE, cum.mnt$RMSE)
[1] 0.06889558 0.07368794
```

À la section 2.4, les CV de l'estimateur \bar{y}_s pour la variable de stratification REV_{84} sont inférieurs à 5 % dans le cas des plans cum \sqrt{f} et géométrique. Pour l'estimation de la moyenne de RM_{T85} , les CV sont de plus de 6 %. On voit bien qu'un calcul des tailles d'échantillon avec une variable de stratification vient sous-estimer le n nécessaire à la réalisation du CV cible pour une variable d'intérêt différente. Ces résultats sont présentés pour les deux premiers plans stratifiés au tableau 8, lequel indique également le plan optimal obtenu en appliquant l'algorithme de Kozak à la variable REV_{84} avec $Y = X$.
Tout comme Rivest (2002), nous ajustons un modèle loglinéaire pour la relation entre les deux variables. Comme on peut le voir à la figure 2, il existe des valeurs extrêmes et le code R qui suit permet d'estimer les paramètres du modèle loglinéaire en écartant les municipalités pour lesquelles le rapport X/Y est atypique. Les 18 municipalités mises de côté sont représentées par un astérisque à la figure 2. On ajuste ensuite un modèle de régression aux données restantes à l'aide de la fonction `lm`.

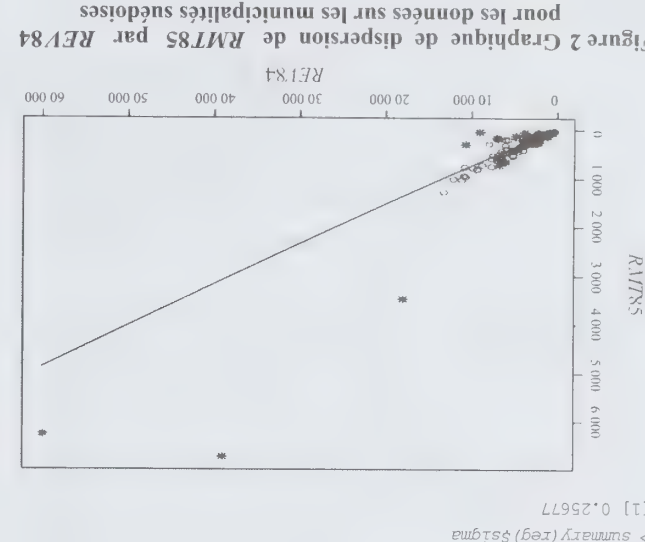


Figure 2 Graphique de dispersion de RM_{T85} par REV_{84} pour les données sur les municipalités suédoises

Le code qui suit stratifie la population MU_{284} par REV_{84} à l'aide des méthodes cum \sqrt{f} et géométrique. On procède toutefois à l'allocation de régression loglinéaire de calculées à l'aide du modèle de régression loglinéaire de RM_{T85} sur REV_{84} . Dans ces deux plans, les strates sont les mêmes que celles qui ont été calculées précédemment.

une recherche aléatoire où on choisit les $L - 1$ limites de strates parmi les valeurs ordonnées de X , avec élimination des doublons. Dans une itération, on prend au hasard une valeur d dans l'ensemble $\{-\text{maxstep}, -\text{maxstep}+1, \dots, \text{maxstep}\}$ et une des $L - 1$ bornes, puis on déplace la limite choisie de d positions dans le vecteur des valeurs triées de X . Si (1) diminué avec la nouvelle borne, cette dernière est conservée, sinon elle est mise de côté et les bornes sont inchangées pour cette itération. L'algorithme s'arrête lorsque les bornes n'ont pas changé pour maxstl1 itérations consécutives. Les valeurs par défaut sont $\text{maxstep}=3$ et $\text{maxstl1}=100$. Deux exécutions consécutives de l'algorithme de Kozak pourraient donner des plans différents, parce que cet algorithme est aléatoire. La fonction `strata.LH` exécute l'algorithme `rep` fois et l'information de chaque itération est contenue dans l'élément `rep.deta11` des objets `R` de classe `strata`. La valeur par défaut est `rep=3`. Si les `rep` exécutions mènent à des plans différents, on peut modifier les paramètres de réglage de l'algorithme. On peut aussi employer `rep="change"`, qui exécute l'algorithme 27 fois avec des valeurs initiales et des `maxstep` différentes. On trouvera en annexe un autre exemple illustrant le cas où l'algorithme de Kozak n'atteint pas un minimum global.

Avec N^n valeurs uniques de X , il y a approximativement $\binom{L-1}{N^n}$ ensembles possibles de limites de strates. Si ce nombre est inférieur à minsol , tous les ensembles possibles de bornes sont testés par opposition à un traitement aléatoire. La valeur par défaut est `minsol=1000`. Les éléments `maxstep`, `maxstl1`, `minsol` et `rep` appartiennent à l'argument `algo.control`. Au tableau 7, nous avons été incapables d'améliorer le plan stratifié issu de la méthode géométrique pour la population `UScites`. Nous présentons ci-après la commande pour exécuter l'algorithme de Kozak 27 fois avec divers paramètres de réglage.

```
> data(UScites)
> pop2LHrep <- strata.LH(x = UScites, CV = 0.0145, Ls = 5,
  alloc = c(0.5, 0, 0.5), takeall = 0, algo = "Kozak",
  algo.control = list(rep = "change"))
```

Cette commande s'exécute en quelques secondes et donne un plan stratifié avec $n = 100$, qui ressemble à ce que présente le tableau 1 pour la population `UScites`.

3.4 Plans avec taille d'échantillon préalable n

Avec l'algorithme de Kozak, il est possible de trouver les bornes qui minimisent le CV de \bar{y}_s pour une taille fixe d'échantillon n plutôt que de minimiser n pour un CV préalable. À titre d'exemple, reprenons les plans stratifiés du tableau 1. Les bornes de la méthode géométrique nous servent de valeurs initiales et nous exécutons l'algorithme de Kozak par défaut. Nous présentons ci-après le code `R` pour la population `Debtors`.

```
> pop1k <- strata.LH(x = Debtors, initbh = pop1$bh, n = 100,
  Ls = 5, alloc = c(0.5, 0, 0.5), algo = "Kozak")
```

Les CV de l'estimateur \bar{y}_s avec les plans stratifiés optimaux sont de 3,12 %, 1,43 %, 1,72 % et 1,04 % respectivement pour les quatre populations comparativement à des valeurs de 3,59 %, 1,45 %, 1,83 % et 1,07 % au tableau 1. Ainsi, l'algorithme itératif permet de réduire les CV.

4. Stratification avec moments anticipés

On peut tenir compte d'une différence entre la variable de stratification X et la variable d'enquête Y en créant un modèle pour la distribution conditionnelle de Y étant donné X . Dans le programme *stratification*, il existe un modèle loglinéaire où

$$Y = \exp(\alpha) X^{\beta} \exp(\alpha \epsilon),$$

ainsi qu'un modèle linéaire hétéroscédastique avec

$$Y = \alpha + \beta X + \sigma \epsilon X^{\gamma}, \quad (2)$$

où α , β , et γ sont des paramètres réels spécifiés par l'utilisateur et ϵ est une variable aléatoire $N(0, 1)$. Un modèle de remplacement aléatoire (Rivest 1999) est également disponible, et il est possible d'ajouter des taux de mortalité par strate (Baillargeon, Rivest et Ferland 2007) au modèle loglinéaire.

Dans ces modèles, la moyenne anticipée de Y pour les unités classées dans la strate h avec $X \in [b_{h-1}, b_h)$ est

$$\bar{Y}_h = \frac{1}{N_h} \sum_{b_{h-1} \leq X_i < b_h} E(Y | X_i)$$

alors que la variance anticipée est

$$S_{Yh}^2 = \frac{1}{N_h} \sum_{b_{h-1} \leq X_i < b_h} \{E(Y | X_i) - \bar{E}(Y | X)\}_i^2 + \frac{1}{N_h} \sum_{b_{h-1} \leq X_i < b_h} \text{Var}(Y | X_i)$$

où $\bar{E}(Y | X)_h$ est la moyenne des valeurs prédites de Y pour les unités de la strate h . Dans les fonctions `strata.cumprobt`, `strata.gmo` et `strata.bh`, ces expressions servent à évaluer les propriétés échantillonnelles de \bar{y}_s , alors que, dans la fonction `strata.LH`, la minimisation de (1) se fait avec les moments anticipés. Dans `strata.LH`, les bornes des strates dépendent du modèle pour la relation entre X et Y ; ce n'est pas le cas pour les autres fonctions `strata`.

4.1 Exemple avec les municipalités suédoises `MT284`

À la section 2.5, nous avons construit deux plans stratifiés pour la population `MT284` avec `REV84` comme variable de stratification. Le code `R` qui suit permet d'examiner la performance de ces plans pour la variable `REV85`. Le vecteur `ord` contient la position des statistiques d'ordre de la variable `REV84`. Ainsi, `Y[ord]` est le vecteur de la variable `REV85` ordonné selon `REV84`.

Au tableau 5, la concordance est bonne entre les bornes présentées au tableau 8 de Sethi (1963) et celles obtenues par simulation. On pourrait employer la même technique pour calculer les bornes optimales des strates à partir d'une distribution arbitraire comme dans Khan, Nand et Ahmad (2008).

Tableau 5
Comparaison des bornes optimales de Sethi (1963) et des bornes approximatives obtenues avec stratification

L	Résultats de stratification				Résultats de Sethi			
	1	2	3	4	1	2	3	4
h_{ij}	2	-0,007	-0,531	0,567	-0,55	0,55	-0,88	0,88
$N(0,1)$	4	-0,883	-0,008	0,864	-1,11	-0,34	0,34	1,11
b_{ij}	2	30,674	26,535	35,141	24,0	30,6	26,0	35,0
X_{30}^2	4	24,340	30,733	38,179	22,0	28,0	33,0	40,0
	5	22,821	28,123	33,386	40,202			

3.2 Exemple tiré de Gunning et Horgan (2004)

Les plans stratifiés construits par Lavallée et Hidiroglou (1988) ont toujours une strate à tirage complet pour une variable d'enquête asymétrique. Pour montrer que la chose n'était pas toujours obligatoire, Gunning et Horgan (2004) ont construit leurs plans stratifiés optimaux avec une strate à tirage complet pour les quatre populations au tableau 1. Les résultats de leur tableau 7 (avec de légères corrections tenant à des erreurs d'arrondis) sont reproduits au tableau 6. Si on compare les tableaux 1 et 6, on constate que les plans optimaux avec strate à tirage complet présentent des valeurs de plus de 100 pour trois populations sur quatre. Le plan optimal est supérieur à celui de la méthode géométrique seulement pour la population Debtors. Nous présentons ci-après le code R pour exécuter l'algorithme de Sethi sur la population Debtors.

```
popLIH <- strata.LH(x = Debtors, CV = 0.0359, Ls = 5,
  alloc = c(0.5, 0, 0.5), takeall = 1, algo = "sethi")
```

Au tableau 6, on s'attendrait à ce que les plans optimaux obtenus à l'aide d'un algorithme itératif aient une taille d'échantillon inférieure à celle des plans géométriques, attente déçue pour trois populations. Ce pourrait être que l'algorithme de Sethi est incapable de trouver la véritable valeur minimale de n . Pour le vérifier, nous avons refait les calculs avec l'argument algo="Kozak". Les tailles d'échantillon n figurent à la deuxième colonne du tableau 7. Avec l'algorithme de Kozak, on trouve une valeur n inférieure à celle de l'algorithme de Sethi pour trois des quatre populations. Cela montre bien la faiblesse qu'accuse l'algorithme de Sethi dans le traitement de populations réelles. À la deuxième colonne du tableau 7, les valeurs n sont de

3.3 Réglage des algorithmes

Population	algo=Sethi	algo=Kozak	takeall=1	takeall=0
Debtors	93	92	104	88
UScities	137	114	107	95
UScolleges	107	107	107	123
USbanks	104	88	104	88

Tableau 7
Taille d'échantillon n pour trois plans optimaux et quatre populations

Population	n	CV	1	2	3	4	5
Debtors	93	0.0359	349,33	1 190,16	3 482,98	10 322,50	146 26
UScities	137	0.0145	1 856	991	350	146 26	20 26
	b_{ij}	14,72	21,62	35,59	80,47		
	N_{ij}	189	270	336	164	79	
UScolleges	107	0.0183	512,32	869,76	1 577,23	3 668,85	30 79
	b_{ij}	133	180	185	110	69	
	N_{ij}	4	8	16	30	79	
USbanks	104	0.0107	99,37	129,60	181,94	317,36	65 74
	b_{ij}	4	6	10	18	69	
	N_{ij}	70	66	82	65	74	
	n_{ij}	4	4	7	15	74	

Tableau 6
Plans stratifiés optimaux avec strate à tirage complet par l'algorithme de Sethi appliqué aux quatre populations du tableau 1

réglent la recherche aléatoire avec algo="Kozak". initiales de strates avec strata.LH et les paramètres qui présenterons maintenant plus en détail le choix de limites Pour mieux comprendre les résultats du tableau 7, nous arguments par défaut qui régissent sa recherche aléatoire. Kozak demeure pire que la méthode géométrique. Il ne trouve pas la véritable valeur minimale de n avec les n . Cependant, pour la population UScities, l'algorithme de strate à tirage complet vient réduire la taille d'échantillon Pour les populations Debtors et UScolleges, le retrait de la Les résultats figurent à la troisième colonne du tableau 7. une strate à tirage complet, c'est-à-dire avec takeall=0. Nous avons réexécuté l'algorithme de Kozak sans exiger parce qu'on n'a pas besoin d'une strate à tirage complet. Le plan de la méthode géométrique pourrait être meilleur, plus de 100 pour deux des quatre populations. Dans ce cas,

Les bornes initiales par défaut pour les deux algorithmes itératifs sont les valeurs initiales arithmétiques de Gunning et Horgan (2007) avec $b_{ij} = \min X + (\max X - \min X) \times h/L$, pour $h = 1, \dots, L - 1$. Au tableau 7, ce choix est contestable et les bornes géométriques auraient été plus proches des bornes optimales. Dans la fonction strata.LH, l'argument initb= permet de spécifier un vecteur de $L - 1$ valeurs initiales des bornes. On peut modifier le nombre maximal d'itérations avec l'élément maxiter de l'argument algo.control.

L'algorithme de Kozak a d'abord été proposé dans Kozak (2004) ; voir aussi Kozak et Verma (2006). On y procède par

stratifié avec strate à tirage complet, la variance de l'estimateur par dilata-tion est donnée par

$$\text{Var}(\bar{y}_s) = \sum_{h=1}^H \left(\frac{N_h}{N} \right)^2 \left(\frac{1}{1 - \frac{N_h}{N}} a_h^2 \right) S_{y_h}^2$$

où $\{a_h\}$ est la règle d'attribution pour le calcul des tailles d'échantillon. Le n qui assure un CV de c est donné par

$$n = N_L + \frac{\sum_{h=1}^H N_h^2 S_{y_h}^2 / (a_h N^2)}{\bar{Y}^2 c^2 + \sum_{h=1}^H N_h S_{y_h}^2 / N^2} \quad (1)$$

Dans cette expression, on peut écrire $n = n(b_1, \dots, b_L)$ pour bien faire voir que la valeur de n dépend des bornes des strates. Avec la fonction `strata.LH`, on vise des limites optimales b_h qui minimisent $n(b_1, \dots, b_{L-1})$. Nous disposons de deux algorithmes de minimisation, soit celui de Sethi (1963) tel que mis en oeuvre par Lavallée et Hidiroglou (1988) avec `algo="Sethi"` et l'algorithme de recherche aléatoire de Kozak (2004) avec `algo="Kozak"`. Le second est l'algorithme par défaut. Dans cette section, $Y = X$; nous ne distinguons donc pas la variable de stratification de la variable d'enquête.

3.1 Exemple tiré de Sethi (1963) avec distribution normale

Un problème classique consiste à déterminer les limites optimales de L strates dans une population infinie à partir d'une distribution connue. Ainsi, Sethi (1963) a calculé des bornes optimales pour les distributions normale et χ^2_{30} . Pour obtenir des solutions approximatives, on peut appliquer la fonction `strata.LH` à une population Monte Carlo simulée à partir de la distribution connue sans demander de strate à tirage complet. Dans (1), on a $N_h/N^2 \approx 0$ et les limites optimales sont les mêmes pour tout CV cible c .

Dans le code R qui suit, nous simulons des populations de taille 10^5 à partir des distributions χ^2_{30} et $N(10, 1)$. Observez que le programme *stratification* exige que la variable de stratification soit positive et, par conséquent, la fonction ne serait pas applicable à une distribution normale standard. En soustrayant 10 des limites $N(10, 1)$ nous obtenons celles de la $N(0, 1)$. Les calculs utilisent la fonction `strata.LH`, avec les arguments `algo="Sethi"` et `takeall=0`, une strate à tirage complet n'est donc pas demandée.

```
> z <- rnorm(100000, 10)
> z15 <- strata.LH(x = z, CV = 0.001, Ls = 5,
+ algo = "Sethi", takeall = 0, algo = "Sethi")
[1] -1.1247340 -0.3480829 0.3297044 1.0979017
+ x158bh
[1] 2.09144e+01 4.00000e+01
```

avant la construction des strates. L'argument `certain` dis-ponible pour les quatre fonctions `strata` rend la chose possible. À titre d'exemple, reprenons la comparaison des plans présentés au tableau 3 pour la méthode de la fonction `cum.f` et la méthode géométrique. On met les trois grandes municipalités de la figure 1 dans une strate à tirage obligatoire et les $N = 281$ autres municipalités dans $Ls=4$ strates par ces deux méthodes de stratification. Nous donnons plus loin le code R pour l'élaboration de ces deux plans. Avec la commande `x=sort(Sweden$REV84)`, on ordonne les municipalités selon *REV84*. Ainsi, les trois grandes municipalités en question sont les éléments 282, 283 et 284 du vecteur trié. Les deux objets R de classe `strata`, à savoir `geo_cer` et `cum_cer`, contiennent chacun un élément `certain.info` qui caractérise la strate à tirage obligatoire.

```
> geo_cer <- strata.geo(x = sort(Sweden$REV84), CV = 0.05,
+ Ls = 4, alloc = c(0.35, 0.35, 0), certain = 282:284)
> cum_cer <- strata.cumcoef(x = sort(Sweden$REV84),
+ nclass = 50, CV = 0.05, Ls = 4, alloc = c(0.35, 0.35, 0),
+ certain = 282:284)
[1] 282 283 284
```

Au tableau 4, le plan `cum.f` est plus efficace que celui de la méthode géométrique. On a intérêt à mettre les trois grandes municipalités dans une strate à tirage obligatoire, car les tailles d'échantillon au tableau 4 sont inférieures à celles du tableau 3. Avec l'argument `certain`, on peut forcer l'inclusion de tout ensemble d'unités dans l'échantillon. On peut s'en servir pour y inclure des unités qui sont extrêmes pour une variable secondaire, différente de la variable de stratification, ou qui se sont déjà révélées hautement instables.

Tableau 4
Deux plans stratifiés visant les municipalités suédoises et élaborés avec une strate à tirage obligatoire

Méthode	1	2	3	4	5	n	CV
géométrique	42	116	88	35	3	24	4,71
N_h	2	5	7	7	3	3	
N_h	127	79	46	29	3	19	4,72
n_h	3	4	4	5	3	3	

3. Méthode d'optimisation

Les méthodes de stratification présentées à la section 2 ne donnent pas toujours un plan stratifié optimal qui minimise la taille d'échantillon n nécessaire à l'obtention du CV cible (ou qui minimise le CV pour un n fixe). Dans cette section, nous présenterons la fonction `strata.LH` pour la détermination de plans optimaux. LH correspond à Lavallée et Hidiroglou (1988), des pionniers de la construction de plans optimaux pour des populations réelles. Dans un plan

cas, la taille d'échantillon non arrondi pour la strate 5 est $n_{adjust5h} \cdot \text{round}[5] = 25,40$ pour $N_5 = 24$ unités. À noter que, lorsque n est grand ou que le CV cible est bas, il peut y avoir plusieurs strates à tirage complet.

Tableau 2

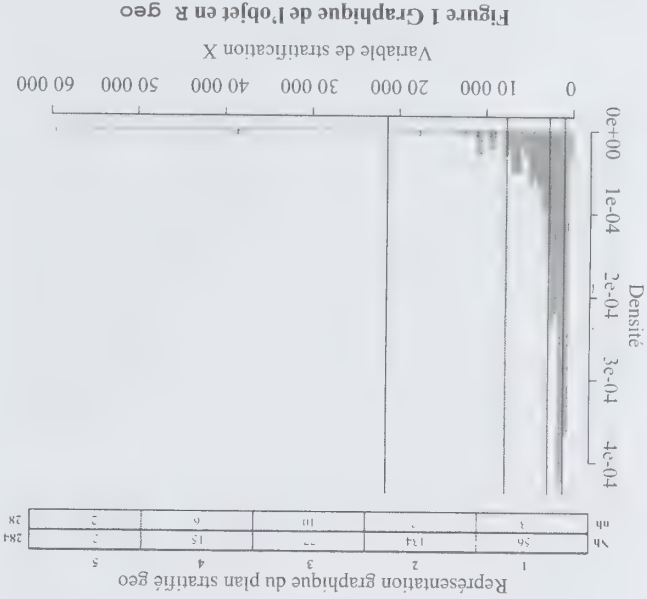
Plans stratifiés obtenus avec et sans correction automatique pour une strate à tirage complet

n	1	2	3	4	5		
						b_h	N_h
correction	100	13	20	25	18	118,59	114
non-correction	99	13	20	24	18	200,92	116

2.4 Ajout d'une strate à tirage complet

Considérons maintenant la base de données de $N = 284$ municipalités suédoises en annexe de Särndal, Swensson et Wretman (1992). Dans les instructions qui suivent, nous employons la méthode géométrique pour répartir cette population en $Ls=5$ strates par la variable REV84, qui donne les valeurs foncières en 1984. Nous employons l'attribution de puissance avec l'exposant 0,7 et $alloc=c(0.35, 0.35, 0)$. L'objet R de classe strata geo contient le plan stratifié. La commande `plot(gco)` produit la figure 1. On y trouve un histogramme de la variable de stratification avec les bornes des strates et un tableau récapitulatif du plan stratifié.

```
> data(Sweden)
> geo <- strata.geo(x = Sweden$REV84, CV = 0.05, Ls = 5,
+ 1100, c(0.35, 0.35, 0))
```



La figure 1 indique que, avec la méthode géométrique de stratification, on se trouve à mettre deux des trois valeurs extrêmes de REV84 dans une strate à tirage complet. Le code R qui suit crée cum, un plan stratifié pour cette

population par la méthode $\text{cum}\sqrt{f}$. Cette méthode $\text{cum}\sqrt{f}$ est d'une application difficile, puisque les classes sont de longueur $\{\max(\text{REV84}) - \min(\text{REV84})\} / 50 = 1191$. On constate à la figure 1 que la plupart des classes présentent une fréquence nulle. En fait, la strate 5 comprend 43 des 50 classes. Dans ce plan, il n'y a pas de strate à tirage complet. Pour calculer les tailles d'échantillon avec une telle strate, on peut employer la fonction `strata.bh` avec les limites $\text{cum}\sqrt{f}$ stockées dans `cum3bh` en prenant la commande `takeall=1`. Nous créons ainsi le troisième plan du tableau 3 cum3 en rendant la taille d'échantillon de la strate 5 du plan $\text{cum}\sqrt{f}$ égale à sa taille de population par la commande `cum3$nh[5]<-cum3$N[5]`. Nous calculons enfin par la fonction `var.strata` la variance de l'estimation \bar{y}_s pour la variable REV84 à l'aide de ce quatrième plan.

```
> cum <- strata.cumrootf(x = Sweden$REV84, nclass = 50,
+ CV = 0.05, Ls = 5, alloc = c(0.35, 0.35, 0))
> cum2 <- strata.bh(x = Sweden$REV84, bh = cum3bh, CV = 0.05,
+ Ls = 5, takeall = 1, alloc = c(0.35, 0.35, 0))
> cum3 <- cum
> cum3$nh[5] <- cum3$N[5]
> cum3.var <- var.strata(cum3, y = Sweden$REV84)
```

Tableau 3
Quatre plans stratifiés de la population de municipalités suédoises

Méthode	1	2	3	4	5	n	CV
N_h	56	134	77	15	2	284	4,83
n_h	3	7	10	6	2	28	4,87
$\text{cum}\sqrt{f}$	120	70	52	27	15	41	4,44
n_h modif1	2	7	9	8	10	24	2,29
n_h modif2	7	2	3	2	15	46	

Il pourrait être bon dans un plan stratifié d'imposer la contrainte de l'échantillonnage d'un certain nombre d'unités

2.5 Strate à tirage obligatoire

Dans `glven` arguments, `model`=none signifie que les propriétés d'échantillonnage de \bar{y}_s , présentées à la fin de la sortie sont évaluées avec $X = X_s$, en l'occurrence pour la variable `loans`. La moyenne est 15,39408 et le CV prévu de 0,0494897 est celui de l'estimateur \bar{y}_s de la moyenne de la variable `loans` obtenue avec ce plan d'échantillonnage. Les limites des strates dans cette sortie sont (10,2, 29,6, 98,5) ; elles sont égales à celles qui figurent au bas de la page 349 dans Cochran (1961), une fois pris en compte les arrondis. Dans `Strata Information`, n_h renvoie aux taux de ré-ponse des strates dont il sera question à la section 5.1. L'objet `cum` en R contient plusieurs éléments énumérés par la commande `names(cum)`.

```
> names(cum)
[1] "n_h"
[2] "n_h"
[3] "n_h.nonint"
[4] "certaln.info"
[5] "varh"
[6] "varh"
[7] "varh"
[8] "varh"
[9] "varh"
[10] "varh"
[11] "varh"
[12] "varh"
[13] "varh"
[14] "varh"
[15] "varh"
[16] "varh"
[17] "varh"
[18] "varh"
[19] "varh"
[20] "varh"
[21] "varh"
[22] "varh"
[23] "varh"
[24] "varh"
[25] "varh"
[26] "varh"
[27] "varh"
[28] "varh"
[29] "varh"
[30] "varh"
[31] "varh"
[32] "varh"
[33] "varh"
[34] "varh"
[35] "varh"
[36] "varh"
[37] "varh"
[38] "varh"
[39] "varh"
[40] "varh"
[41] "varh"
[42] "varh"
[43] "varh"
[44] "varh"
[45] "varh"
[46] "varh"
[47] "varh"
[48] "varh"
[49] "varh"
[50] "varh"
[51] "varh"
[52] "varh"
[53] "varh"
[54] "varh"
[55] "varh"
[56] "varh"
[57] "varh"
[58] "varh"
[59] "varh"
[60] "varh"
[61] "varh"
[62] "varh"
[63] "varh"
[64] "varh"
[65] "varh"
[66] "varh"
[67] "varh"
[68] "varh"
[69] "varh"
[70] "varh"
[71] "varh"
[72] "varh"
[73] "varh"
[74] "varh"
[75] "varh"
[76] "varh"
[77] "varh"
[78] "varh"
[79] "varh"
[80] "varh"
[81] "varh"
[82] "varh"
[83] "varh"
[84] "varh"
[85] "varh"
[86] "varh"
[87] "varh"
[88] "varh"
[89] "varh"
[90] "varh"
[91] "varh"
[92] "varh"
[93] "varh"
[94] "varh"
[95] "varh"
[96] "varh"
[97] "varh"
[98] "varh"
[99] "varh"
[100] "varh"
```

On peut imprimer un élément de l'objet `cum` de classe `strata` en tapant `cum$h` suivi du nom de l'objet. Ainsi, la commande `cum$stratumID` imprime la strate de chaque unité de la population. La variable `cum$nclasssh` est propre à la fonction `strata.cumrootf` et indique comment les `nclass=20` classes initiales ont été réunis en trois strates.

Dans cette stratification, les strates 1, 2 et 3 contiennent respectivement 2, 4 et 14 des `nclass=20` classes initiales.

2.2 Méthode géométrique

La méthode géométrique de stratification a été introduite par Gunning et Horgan (2004). Elle délimite les strates par $b_h = \min X \times (\max X / \min X)^{h/L}$, pour $h = 1, \dots, L - 1$. Une fois les bornes b_h établies, les calculs de taille d'échantillon des strates sont les mêmes qu'avec la méthode `strata.cumrootf`.

Pour illustrer, nous avons stratifié les quatre populations présentées dans Gunning et Horgan (2004), à savoir Debtors, USBanks, USCities et USColleges, en $L=5$ strates. Les trois dernières populations ont été examinées dans Cochran (1961). Les quatre sont stockées dans `strata` (`data(Debtors)`. On appelle la première par la commande `data(Debtors)`. Plutôt que de spécifier un CV cible, nous fixons la taille d'échantillon à $n = 100$ tel que fait par Gunning et Horgan (2004). Les commandes qui suivent créent l'objet R `pop1` qui contient le plan stratifié de la population Debtors.

Le tableau 1 récapitule les plans stratifiés obtenus en appliquant la méthode géométrique aux quatre populations en

2.3 Strate à tirage complet

Population	CV	Plans stratifiés pour quatre populations avec $n = 100$				
		1	2	3	4	5
Debtors	0,0359	b_h 148,28	N_h 549,67	n_h 2 037,60	7 553,33	265 51
		b_h 1 054	N_h 1 267	n_h 732	265 51	
		b_h 3	N_h 14	n_h 27	33 23	
		b_h 18,17	N_h 33,01	n_h 59,98	108,98	
		b_h 364	N_h 418	n_h 130	87 39	
		b_h 18	N_h 28	n_h 17	20 17	
		b_h 434,00	N_h 941,76	n_h 2 043,61	4 434,60	
		b_h 94	N_h 255	n_h 198	74 56	
		b_h 3	N_h 15	n_h 27	20 35	
		b_h 118,59	N_h 200,92	n_h 340,39	576,68	
		b_h 114	N_h 116	n_h 64	39 24	
		b_h 13	N_h 20	n_h 25	18 24	

Tableau 1 Plans stratifiés pour quatre populations avec $n = 100$

suivant le fichier d'assistance.

méthodes d'arrondissement dans *stratification* en con-

de la fonction `strata.geo`.

Pour illustrer ce point, nous employons la fonction `strata.bh` et procédons à l'attribution sans correction pour la strate à tirage complet. Nous répartissons l'échantillon et calculons le degré de précision de \bar{y}_s pour un ensemble préalable de bornes de strates. En réglant `takeall.adjust=FALSE`, nous servons de l'attribution de Neyman dans les cinq strates et, comme $n_s > N_s$, le résultat affiché est $n_s = N_s$. Le code qui suit en R prend les bornes de strates géométriques $\{b_h\}$ dans l'objet de classe `strata.adjust`. La fonction `strata.bh` est ensuite appliquée pour obtenir le plan d'échantillonnage sans correction pour la cinquième strate à tirage complet dans l'objet de classe `strata.noadjust`.

Les deux plans sont présentés au tableau 2. Sans la strate à tirage complet, la taille d'échantillon est $n = 99$. Dans ce

Dans toute cette section, $Y = X$. L'emploi d'une même variable pour la stratification et l'évaluation de la précision des estimations d'enquête pourrait avoir pour conséquence une sous-estimation des variances. Nous traiterons à la section 4 de la question du calcul des variances en cas de non-correspondance $Y \neq X$.

2.1 Méthode \sqrt{f}

Cet algorithme de stratification présenté au chapitre 5A de Cochran (1977) s'exécute par la fonction `strata.cumrootf`. Ses arguments sont `x`, vecteur de la variable de stratification dans la population, `ncl`, nombre de classes de taille égale pour la variable `x`, un CV cible pour \sqrt{f} , ou une taille d'échantillon préétablie `n` avec `ts`, nombre de strates, et `alloc`, règle d'attribution. Cet algorithme réunit les `ncl` classes en `ts` strates de sorte que la somme des racines carrées des fréquences dans les classes soit approximativement égale d'une strate à l'autre.

En guise d'illustration, prenons la proportion de prêts industriels de $N = 13\,435$ banques dans Cochran (1961). Nous stratifions cette population et évaluons la taille d'échantillon à prévoir pour que \sqrt{f} ait un CV de 5 % avec l'allocation de Neyman. Le code R qui suit crée le vecteur de la variable de stratification `loans` donnée au tableau 2 de McEvoy (1956). Nous appliquons ensuite la fonction `strata.cumrootf` à cette variable. Comme dans le tableau 2 de Cochran (1961), trois strates seront créées à partir de 20 classes, donc `ncl` est fixé à 20 et `ts`=3. La sortie est enregistrée dans `cum`, un objet R de classe `strata`. Si on tape `cum` ou `print(cum)` dans la fenêtre de commande R, on imprime le plan d'échantillonnage en détail. Les arguments d'entrée, par défaut ou spécifiés par l'utilisateur, s'affichent en premier. Les informations relatives aux strates s'affichent ensuite : bornes, tailles de strate N_h et tailles d'échantillon n_h . La troisième partie de la sortie renseigne sur les propriétés échantillonales de \sqrt{f} .

```
> values <- c(seq(0.5, 9.5, 1), seq(12.5, 97.5, 5))
> nrep <- c(1985, 261, 339, 405, 474, 478, 506, 569, 464, 499,
157, 1581, 1142, 746, 512, 376, 265, 207, 126, 177, 87, 50,
39, 25, 16, 19, 2, 3)
> loans <- rep(values, nrep)
> cum <- strata.cumrootf(x = loans, ncl = 20, CV = 0.05,
ts = 3, alloc = c(0.5, 0, 0.5))
> cum
Given arguments:
model = none
allocation : q1 = 0.1, q2 = 0, q3 = 0.5
ncl = 20, CV = 0.05, ts = 3
x = loans
Total sample size: 50
Anticipated population mean: 15.39408
Anticipated CV: 0.0494897
```

```
Strata information:
  strata  |  bh  |  antcip | Mean antcip | var  |  Nh  |  nh  |  fh
Stratum 1 | 10.2 | 4.12   | 10.46       | 5980 | 14   | 0.00
Stratum 2 | 79.6 | 1.92   | 7.74        | 5626 | 70   | 0.10
Stratum 3 | 98.5 | 44.47  | 165.83      | 1829 | 16   | 0.01
Total      |      | 44.47  | 165.83      | 13435 | 50   | 0.00
```

bornes géométriques. Quant à `strata.LH`, elle construit des plans stratifiés optimaux à l'aide d'algorithmes itératifs. La dernière fonction traite des bornes de strates fournies par l'utilisateur. Avec ces quatre fonctions, on construit des strates, établit des tailles d'échantillon de strates et calcule le degré de précision de l'estimateur simple par dilataion \sqrt{f} de \bar{Y} , la moyenne de la variable d'intérêt Y dans la population qui est liée à la variable de stratification X .

Dans les quatre fonctions `strata`, on emploie la règle de Hidiroglou et Srinath (1993) pour attribuer aux strates les n unités de l'échantillon. Les tailles d'échantillon des strates sont proportionnelles à $N_h^h \frac{F_{2q_3}^h}{S_{2q_3}^h}$, où N_h est la taille de la strate h , et où $\frac{F_{2q_3}^h}{S_{2q_3}^h}$ et $\frac{F_{2q_3}^h}{S_{2q_3}^h}$ sont respectivement la moyenne et la variance anticipées de Y dans cette strate. Dans les fonctions `strata`, on se trouve à spécifier une règle d'allocation par l'argument `alloc` qui contient les exposants (q_1, q_2, q_3) . L'attribution de Neyman correspond à `alloc=c(1/2, 0, 1/2)`. Dans une fonction `strata`, on prend comme entrées le vecteur de population de la variable de stratification `X`, le nombre de strates `ts` et une taille totale d'échantillon `n` ou un CV cible pour l'estimateur simple par dilataion \sqrt{f} . La sortie est un objet R de classe `strata` qui définit un plan stratifié. On y trouve un ensemble de strates déterminées par leurs bornes supérieures $\{b_h\}$ et les tailles de population et d'échantillon de strates N_h et n_h . Dans le programme *stratification*, il existe une cinquième fonction appelée `var.strata` ayant pour entrées un objet R de classe `strata` et un vecteur donnant les valeurs d'une variable d'intérêt Y dans la population. Sa sortie est la variance de \sqrt{f} pour la variable Y et le plan stratifié donné en entrée.

Le texte comporte des instructions en R à taper dans une fenêtre de commande. Ces lignes commencent par `>`. Le texte présente aussi des sorties telles qu'elles s'impriment dans une fenêtre de commande. Une police spéciale permet de reconnaître facilement les commandes et les sorties R, qu'elles se retrouvent dans un paragraphe discutant ou encore dans le texte. L'annexe offre un tableau récapitulatif énumérant tous les arguments possibles des cinq fonctions de *stratification*. Lorsqu'on utilise ce programme, l'instruction `help(stratification)` ouvre un fichier d'assistance cliquable qui renseigne en détail sur le programme et donne des exemples pouvant être collés dans une fenêtre de commande.

2. Méthodes de stratification de base

Dans cette section, nous examinerons deux méthodes élémentaires de stratification, soit la méthode \sqrt{f} de Dalenius et Hodges (1959) et la méthode géométrique de Gunning et Horgan (2004). Ces méthodes sont à calcul exact et ne font donc pas intervenir d'algorithme itératif.

Élaboration de plans stratifiés en R à l'aide du programme *stratification*

Sophie Baillargeon et Louis-Paul Rivest¹

Résumé

Ce document présente un programme R pour la stratification d'une population d'enquête à l'aide d'une variable unidimensionnelle X et pour le calcul de tailles d'échantillon dans les strates. Nous y employons des méthodes non itératives pour délimiter les strates, comme la méthode de la racine carrée des fréquences et la méthode géométrique. Nous pouvons élaborer des plans optimaux où les bornes de strates minimisent soit le CV de l'estimateur simple par dilution pour une taille fixe d'échantillon n , soit la valeur n pour un CV fixe. Nous disposons de deux algorithmes itératifs pour le calcul des bornes optimales. Le plan peut comporter des strates à tirage obligatoire qui sont définies par l'utilisateur et dont toutes les unités sont échantillonnées. Il est également possible d'inclure dans le plan stratifié des strates à tirage complet et à tirage nul qui permettent souvent de réduire les tailles d'échantillon. Les calculs de taille d'échantillon sont fondés sur les moments anticipés de la variable d'enquête X étant donné la variable de stratification X . Le programme traite les distributions conditionnelles de X étant donné X qui sont soit un modèle linéaire hétéroscédastique soit un modèle logarithmique. Nous pouvons tenir compte de la non-réponse par strate dans l'élaboration du plan d'échantillonnage et dans les calculs de taille d'échantillon.

Mots clés : Modèles linéaires ; modèles logarithmiques ; stratification optimale ; échantillonnage d'enquête ; strate à tirage complet ; strate à tirage nul.

1. Introduction

L'établissement de strates et l'élaboration d'un plan stratifié sont une question importante en méthodologie d'enquête depuis les travaux de pionnier de Dalenius il y a plus de soixante ans. Il sera question ici d'une stratification à l'aide d'une variable unidimensionnelle positive X connue pour toutes les unités de la population. On pose que X est en relation avec la variable d'enquête Y . La strate h contient toutes les unités ayant une valeur X dans l'intervalle $[b_{h-1}, b_h)$ pour $h = 1, \dots, L$, de sorte que $b_0 = \min X$ et $b_L = \max X + 1$, où $\min X$ et $\max X$ sont respectivement les valeurs minimale et maximale de la variable de stratification.

La délimitation optimale de strates a une longue histoire ; on peut consulter à ce sujet le chapitre 5A dans Cochran (1977). La méthode de la fonction cumulative de la racine carrée des fréquences (cum \sqrt{f}) de Dalenius et Hodges (1959) apporte une solution approximative à ce problème. Les cas où X a une distribution asymétrique sont fréquents dans les enquêtes auprès des entreprises et ont reçu une attention particulière. Guinning et Horgan (2004) ont proposé une méthode géométrique de stratification et Hidroglou (1986) a fait valoir que les grandes unités devraient être remises dans une strate à tirage complet. Plutôt que de s'en remettre à une méthode approximative pour la construction des strates, Lavallée et Hidroglou (1988) ont proposé un algorithme itératif qui donne des bornes optimales pour une variable X déterminée. Parfois cet algorithme ne converge pas (Detlefsen et Veum 1991), Slanta et Krenzke (1996) ont

Dans cet article, nous présentons le programme R *stratification* qui met en oeuvre la plupart des méthodes que nous venons de mentionner. Il s'agit d'un environnement convivial pour l'élaboration de plans stratifiés et l'évaluation de leur rendement pour des populations réelles. Nous présentons ce programme en revisitant des exemples déjà traités dans la littérature qui illustrent ses caractéristiques importantes. Les quatre fonctions de *stratification* dont le nom débute par *strata* construisent des plans d'échantillonnage stratifiés. Ce sont les fonctions *strata.cumrootf*, *strata.geo*, *strata.lh* et *strata.bh*. Les deux premières correspondent aux méthodes simples du cum \sqrt{f} et des

démontre que, dans certains cas, les bornes optimales ne sont pas définies d'une manière unique. On a proposé d'autres méthodes comme l'algorithme de recherche aléatoire de Kozak (2004) pour parer à certaines de ces difficultés. L'hypothèse d'une correspondance exacte entre la variable d'enquête Y et la variable de stratification X est irréaliste lorsqu'on calcule des tailles d'échantillon et plusieurs auteurs, dont Dayal (1985) et Sigman et Monsoon (1995), ont voulu répartir l'échantillon entre les strates en se fondant sur les moments anticipés de Y si on sait que X se trouve dans $[b_{h-1}, b_h)$. Sweet et Sigman (1995) et Rivest (1999, 2002) ont suggéré de faire entrer ces moments anticipés dans l'algorithme de stratification de Lavallée et Hidroglou (1988). Récemment, Baillargeon et Rivest (2009) ont montré que, en mettant les petites unités dans une strate à tirage nul, c'est-à-dire hors échantillonnage, on pouvait réduire la taille d'échantillon nécessaire à l'obtention d'un degré de précision préalable.

1. Sophie Baillargeon, Département de mathématiques et de statistique, 1045, avenue de la Médecine, Université Laval, Québec, Canada G1V 0A6. Courriel : sophie.baillargeon@mat.ulaval.ca ; Louis-Paul Rivest, Département de mathématiques et de statistique, 1045, avenue de la Médecine, Université Laval, Québec, Canada G1V 0A6. Courriel : louis-paul.rivest@mat.ulaval.ca.

- Chambers, R., et Tzavidis, N. (2006). M-quantile models for small area estimation. *Biometrika*, 93, 255-268.
- Chandra, H., et Chambers, R.L. (2005). Comparing EBLUP and C-EBLUP for small area estimation. *Statistics in Transition*, 7, 637-648.
- Chandra, H., et Chambers, R. (2009). Multipurpose weighting for small area estimation. *Journal of Official Statistics*, 25(3), 379-395.
- Chandra, H., Salvati, N., et Chambers, R. (2007) Small area estimation for spatially correlated populations. A comparison of direct and indirect model-based methods. *Statistics in Transition*, 8, 887-906.
- Chen, G., et Chen, J. (1996). Une méthode de transformation applicable à l'échantillonnage de population finies calés par une méthode de vraisemblance empirique. *Techniques d'enquête*, 22, 139-147.
- Harville, D.A. (1977). Maximum likelihood approaches to variance component estimation and to related problems. *Journal of the American Statistical Association*, 72, 320-338.
- Hidiroglou, M.A., et Smith, P.A. (2005). Developing small area estimates for business surveys at the ONS. *Statistics in Transition*, 7, 527-539.
- Jiang, J., et Lahiri, P. (2006). Estimation of finite population domain means: A model-assisted empirical best prediction approach. *Journal of the American Statistical Association*, 101, 301-311.
- Karberg, F. (2000a). Population total prediction under a lognormal superpopulation model. *Metron*, LVIII, 53-80.
- Karberg, F. (2000b). Survey estimation for highly skewed populations in the presence of zeroes. *Journal of Official Statistics*, 16, 229-241.
- Longford, N.T. (2007). De l'erreur-type des estimateurs pour petits domaines fondés sur un modèle. *Techniques d'enquête*, 33, 81-92.
- McCulloch, C.E., et Searle, S.R. (2001). *Generalized, Linear and Mixed Models*. New York : John Wiley & Sons, Inc.
- Prasad, N.G.N., et Rao, J.N.K. (1990). The estimation of the mean squared error of small area estimators. *Journal of the American Statistical Association*, 85, 163-171.
- Rao, J.N.K. (2003). *Small Area Estimation*. New York : John Wiley & Sons, Inc.
- Royal, R.M. (1976). The linear least-squares prediction approach to two-stage sampling. *Journal of the American Statistical Association*, 71, 657-664.
- Royal, R.M., et Cumberland, W.G. (1978). Variance estimation in finite population sampling. *Journal of the American Statistical Association*, 73, 351-358.
- Slud, E. V., et Maiti, T. (2006). Mean squared error estimation in transformed Fay-Herriot models. *Journal of the Royal Statistical Society, Séries B*, 68(2), 239-257.
- Tzavidis, N., Salvati, N., Pratesi, M. et Chambers, R. (2008). M-quantile models with application to poverty mapping. *Statistical Methods and Applications*, 17, 393-411.
- Valliant, R., Dorfman, A.H. et Royall, R.M. (2000). *Finite Population Sampling and Inference*. New York : John Wiley & Sons, Inc.
- Wu, C., et Sitter, R.R. (2001). A model calibration approach to using complete auxiliary information from survey data. *Journal of the American Statistical Association*, 96, 185-193.
- Bates, D.M., et Pinheiro, J.-C. (1998). Computational Methods for Multilevel Models. <http://franz.stat.wisc.edu/pub/NLME/>.
- Carroll, R., et Ruppert, D. (1988). *Transformation and Weighting in Regression*. New York : Chapman and Hall.
- Statistique Canada, N° 12-001-X au catalogue

Remerciements

Le premier auteur est reconnaissant à la Commonwealth Scholarship Commission du Royaume-Uni du soutien financier que lui a procuré la bourse de doctorat qui lui a été accordée. Les auteurs remercient également le rédacteur en chef, le rédacteur associé et deux examinateurs de leurs commentaires constructifs qui leur ont permis de réviser la version originale de l'article et de l'améliorer considérablement.

Le prédicteur indirect (20) de la moyenne de petit domaine est obtenu en appliquant des notions de prédiction bien connues. Sous les modèles avec transformation logarithmique, diverses approches existent pour obtenir le meilleur prédicteur indirect de la moyenne de petit domaine. Par exemple, Slud et Maiti (2006) ont décrit un prédicteur indirect de la moyenne de petit domaine sous une version au niveau du domaine du modèle logarithmiquement transformé (9). Berg (2009, communication privée) suit l'approche de Slud-Maiti pour obtenir un prédicteur de la moyenne de petit domaine sous une spécification avec ordonnée aléatoire du modèle avec transformation logarithmique (9) au niveau de l'unité. Cependant, comme celui de Slud-Maiti, le prédicteur de Berg ne tient pas compte de la correction du biais qui est nécessaire après la rétrotransformation pour revenir à l'échelle originale. Les propriétés empiriques de ce prédicteur n'ont pas encore été examinées.

Bibliographie

7. Conclusion et travaux à venir

Les résultats des simulations exposées à la section précédente montrent que le fait de combiner des poids calés sur un modèle fondés sur un modèle avec l'estimation directe peut accroître considérablement l'efficacité de l'estimation sur petits domaines si la relation entre les données de population est clairement non linéaire. Comme il faut s'y attendre, ces gains sont moins importants quand le modèle non linéaire supposé est mal spécifié. Bien que nous ne donnions pas les détails, nos conclusions n'ont essentiellement pas changé quand nous avons exécuté des simulations similaires en utilisant des effets aléatoires suivant une loi gamma.

Notre mise en garde la plus importante en ce qui concerne l'utilisation des poids calés sur un modèle fondés sur un modèle (17) pour l'estimation sur petits domaines tient à leur spécificité. Ces poids ne semblent pas posséder les mêmes caractéristiques « polyvalentes » que les poids EBLUP classiques pour le calcul du total fondés sur des modèles linéaires mixtes. D'autres travaux seront donc nécessaires afin de déterminer comment construire, pour l'estimation sur petits domaines, des poids calés sur un modèle qui sont plus « polyvalents ». Nous nous attendons à ce que de tels poids ne soient pas aussi efficaces que les poids propres aux variables (17), mais nous espérons que cela sera plus que compensé par leur plus grande utilité. Un autre problème très important en pratique est que les variables d'enquête dont la distribution est positivement asymétrique peuvent également prendre des valeurs nulles (voire même négatives). Par exemple, les variables économiques, telles que le passif et les dépenses en immobilisations, prennent souvent des valeurs nulles, tandis que les variables définies comme étant la différence entre deux quantités non négatives (par exemple les bénéfices, qui correspondent à la différence entre les revenus et les dépenses) peuvent être négatives. Karlberg (2000b) utilise un mélange de modèles pour caractériser les données qui comprennent un mélange de valeurs nulles et de valeurs strictement positives. Ce genre de modèle peut être utilisé dans la pondération calée sur un modèle fondée sur un modèle. Enfin, nous notons que l'utilisation d'une approche DFM fondée sur une transformation dans laquelle les hypothèses habituelles concernant le modèle linéaire ne sont qu'approximativement valides (la situation considérée dans le présent article) n'est pas la seule qui a été proposée pour résoudre ce problème. Deux autres approches décrites dans la littérature sont celles du pseudo-EBLUP (Rao 2003, section 7.2.7) et de l'estimateur de type bayésien empirique (BE) assisté par un modèle de Jiang et Lahiri (2006). Rappelons que, selon (8), le prédicteur EBLUP est défini en remplaçant la moyenne de domaine i m_{ij} inconnue par une

$$m_{ij}^w = \left(\sum_{j \in s_i} \pi_{ij}^{-1} \right)^{-1} \sum_{j \in s_i} \pi_{ij}^{-1} y_{ij} = \sum_{j \in s_i} w_{ij} y_{ij} \quad (21)$$

et les valeurs de X dans le domaine i . Autrement dit, sous (3), le pseudo-EBLUP de m_{ij} est

m_{ij}^w pseudo EBLUP

$$= E\{m_{ij}^w | m_{ij}^w, \mathbf{x}_{ij}, \mathbf{x}_{ij}^w\} \\ = \mathbf{x}_{ij}^w \hat{\beta}_w + (\hat{\mathbf{g}}_{ij}^w \hat{\Sigma}_{w, w}^{-1} \hat{\mathbf{g}}_{ij}^w)$$

$$\left(\hat{\mathbf{g}}_{ij}^w \hat{\Sigma}_{w, w}^{-1} \hat{\mathbf{g}}_{ij}^w + \hat{\sigma}_w^2 \sum_{j \in s_i} \pi_{ij}^{-1} \right)^{-1} (m_{ij}^w - \mathbf{x}_{ij}^w \hat{\beta}_w) \quad (22)$$

où $\hat{\beta}_w$, $\hat{\Sigma}_{w, w}$ et $\hat{\sigma}_w^2$ sont les estimations du pseudo-maximum de vraisemblance basées sur les poids w_{ij} , et $\hat{\mathbf{g}}_{ij}^w$ et \mathbf{x}_{ij}^w sont les estimations convergentes sous le plan de \mathbf{g}_{ij} et \mathbf{x}_{ij} qui sont définies exactement de la même manière que m_{ij}^w susmentionnée. Sous le même modèle, l'approche de type BE assistée par modèle de Jiang et Lahiri (2006) mène à un estimateur qui est également défini par conditionnement sur la valeur de m_{ij}^w .

$$m_{ij}^w = \sum_{j \in s_i} w_{ij} E\{E(y_{ij} | \mathbf{x}_{ij}, \mathbf{u}_i) | m_{ij}^w, \mathbf{x}_{ij}\} \\ = \mathbf{x}_{ij}^w \hat{\beta}_w + \{\mathbf{w}_{ij}^w \hat{\Sigma}_{w, w}^{-1} \mathbf{g}_{ij}^w + \hat{\sigma}_w^2 \mathbf{I}_{s_i}\}^{-1}$$

$$\{\mathbf{w}_{ij}^w \hat{\Sigma}_{w, w}^{-1} \mathbf{g}_{ij}^w\} (m_{ij}^w - \mathbf{x}_{ij}^w \hat{\beta}_w) \quad (23)$$

où \mathbf{w}_{ij}^w est le vecteur des poids d'échantillon normalisés dans le domaine i . Notons que, dans (23), nous utilisons des estimations optimales (c'est-à-dire MV ou MVR) pour les paramètres du modèle.

Les estimateurs (22) et (23) sont essentiellement motivés par l'idée d'estimer la moyenne du domaine i par son espérance conditionnelle sous (3), sachant la valeur de l'estimateur convergent sous le plan habituel (21) pour cette quantité. Il s'agit donc d'estimateurs indirects tels que l'estimateur HT-EBLUP. Sous (3), ni l'un ni l'autre n'est aussi efficace que l'estimateur HT-EBLUP, tandis que si (9) plait que (3) est vérifié, les deux estimateurs tirent parti de la convergence sous le plan de m_{ij}^w pour leur robustesse. Puisque s'appuyer sur une propriété en grand échantillon d'une statistique sur petit échantillon semble assez optimiste, nous préférons nous attaquer directement au problème de spécification du modèle, en remplaçant (3) par (9)

est essentiellement dû à une région (21) de l'échantillon original de l'AAGIS qui contenait une valeur fortement aberrante (CDT > 30 000 000 \$A). Cette valeur aberrante a été incluse dans la population utilisée pour la simulation (deux fois) puis sélectionnée (dans un cas, deux fois) dans 37 des 1 000 échantillons de simulation, ce qui a donné lieu à des estimations entièrement irréalistes produites par HT-DFMTr et par HJ-DFMLin pour la région 21. La colonne de droite du tableau 4 donne par conséquent les mesures de performance moyennes de diverses méthodes quand cette région est exclue. Nous voyons ainsi que HT-DFMTr et HJ-DFMLin sont essentiellement à égalité, dominant tous deux HT-EBLUPLin. La raison pour laquelle HT-DFMTr ne fournit pas de gains significatifs par rapport à HJ-DFMLin dans ce cas tient au fait que les modèles linéaires mixtes sur l'échelle originale et sur l'échelle logarithmique sont tous deux relativement mal ajustés aux données de l'AAGIS.

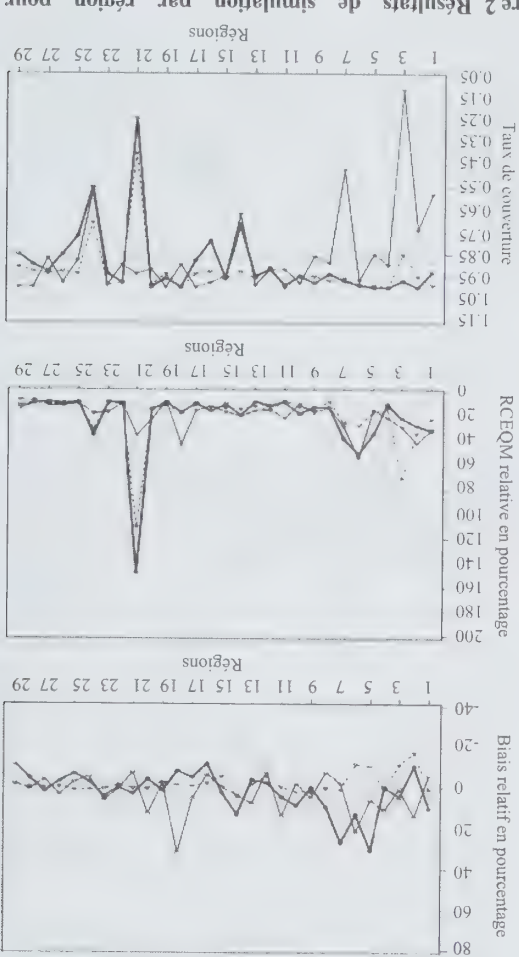


Figure 2 Résultats de simulation par région pour HT-DFMTr (trait épais, 0), HT-EBLUPLin (trait fin, Δ) et HJ-DFMLin (trait pointillé, Δ) dans les simulations fondées sur le plan de sondage basé sur les données de l'AAGIS. Les graphiques représentent (de haut en bas) BR (%), RCEQM relative (%) et TC. Les régions sont classées par ordre croissant de taille de population

Dans l'ensemble, les résultats montrent que, quand le modèle de population sous-jacent n'est pas linéaire, l'utilisation des estimateurs DFM de type HT pour calculer les moyennes de petit domaine fondées sur les poids calés sur un modèle (17) produit des gains importants comparativement aux estimateurs fondés sur un modèle linéaire mixte classiques, tels que HJ-DFMLin et HT-EBLUPLin. Les résultats montrent aussi que l'estimateur indirect HT-EBLUPLin donne de relativement meilleurs résultats que l'estimateur direct HJ-DFMLin dans ces situations. Le prédicteur indirect PETr fondé sur le modèle à transformation logarithmique (9) a de bonnes propriétés en ce qui concerne le biais relatif, mais est moins efficace que l'estimateur DFM sous le même modèle.

Pour l'ensemble B de simulations fondées sur un modèle, nous avons étudié la robustesse de l'estimation directe calée sur un modèle fondé sur un modèle à l'erreur de spécification du modèle non linéaire. Les résultats du tableau 3 montrent que, dans ce cas, le biais généré par HT-DFMTr augmente à mesure que le modèle non linéaire réel s'écarte du modèle non linéaire supposé ($\gamma = 0, 0$ dans le tableau). Cependant, ces biais sont compensés par une faible variabilité, de sorte que si l'on considère la moyenne des RCEQM relative, HT-DFMTr donne encore d'aussi bons. voir de meilleurs, résultats que HT-EBLUPLin et continue d'être supérieur à HJ-DFMLin. Les biais produits par HJ-DFMLin et HT-EBLUPLin sont du même ordre, tandis que la moyenne des RCEQM relative de HT-EBLUPLin est plus grande que celle de HJ-DFMLin. Les taux de couverture moyen de HT-EBLUPLin sont marginalement meilleurs que ceux de HJ-DFMLin et de HT-DFMTr, mais la largeur moyenne des intervalles de confiance qui sous-tendent ces taux la plus petite a tendance à être observée pour HT-DFMTr, suivi par HT-EBLUPLin, puis par HJ-DFMLin. Globalement, les résultats de nos simulations fondées sur un modèle pour l'ensemble B indiquent que, même si l'estimation sur petits domaines au moyen d'un estimateur DFM avec des poids calés sur un modèle fondé sur un modèle peut donner lieu à un biais dû à l'erreur de spécification du modèle, la performance globale de cette approche semble être assez peu affectée par de légers écarts par rapport au modèle non linéaire supposé.

Dans le tableau 4 et à la figure 2, nous présentons les mesures de performance moyennes et par région produites par diverses méthodes d'estimation sur petits domaines pour les données de l'AAGIS, respectivement. Ces résultats montrent que le biais relatif moyen de HT-DFMTr est plus faible que celui de HT-EBLUPLin ainsi que celui de HJ-DFMLin, tandis que la moyenne des RCEQM relative de HT-DFMTr est légèrement supérieure aux valeurs correspondantes pour HJ-DFMLin et HT-EBLUPLin. L'examen de la figure 2 révèle que ce résultat

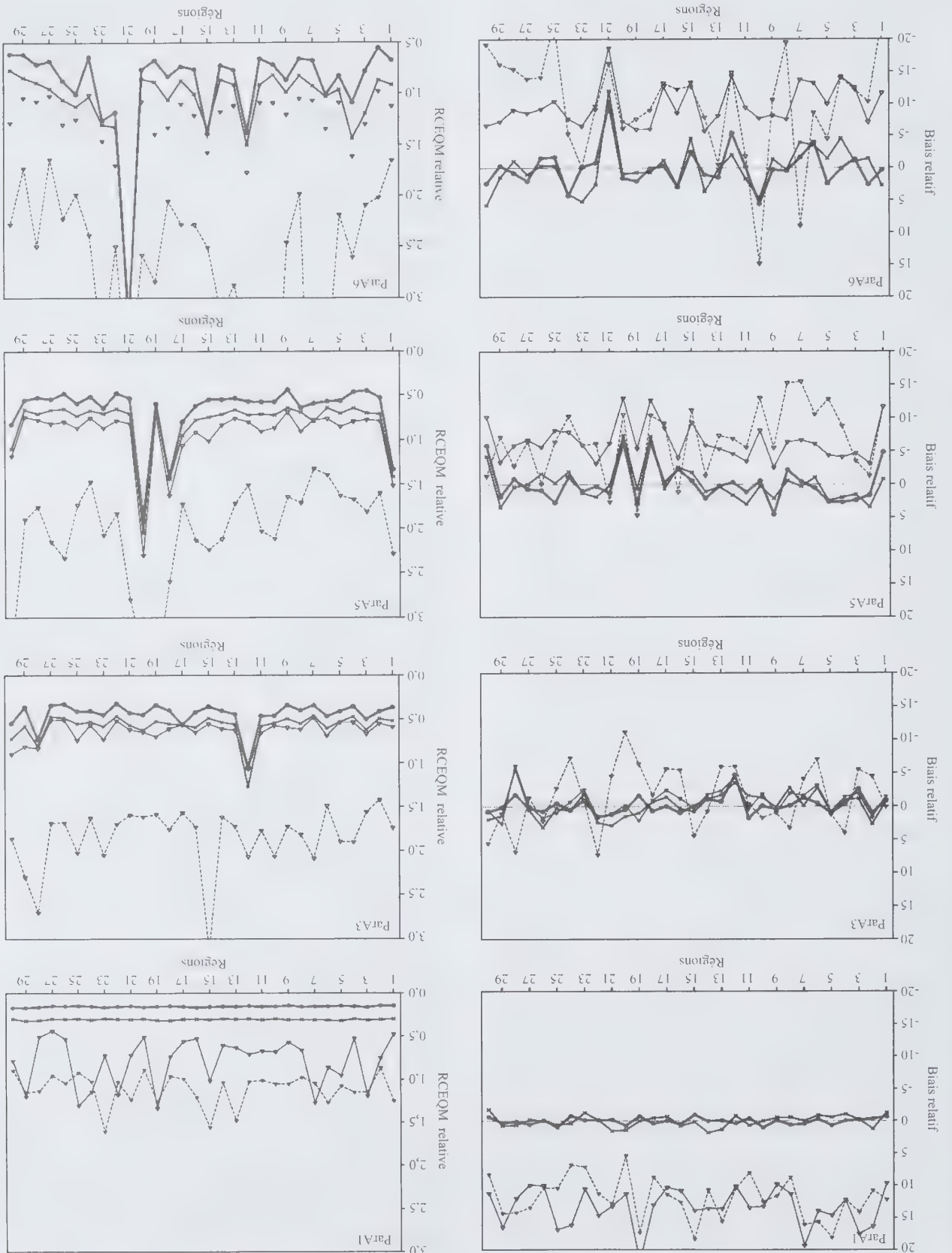


Figure 1 Résultats par région pour HT-DFMTR (trait plein, ●), PETR (trait épais, ×), HT-EBLUPlin (trait fin, Δ) et HT-DFMLin (trait pointillé, Δ) sous les ensembles de paramètres 1 (ParA1), 3 (ParA3), 5 (ParA5) et 6 (ParA6). La colonne de gauche donne le biais relatif (%) et la colonne de droite, la RCEQM relative

A de l'étude par simulation. À partir de maintenant, nous concentrons par conséquent notre discussion sur les quatre estimateurs HT-DFMTr, PETr, HJ-DFMLin et HT-EBLUPLin.

Le tableau 2 montre que les biais relatifs moyens et les moyennes des RCEQM relatives pour HT-DFMTr sont systématiquement inférieures à ceux produits par HJ-DFMLin et HT-EBLUPLin. Le biais relatif moyen de HT-DFMTr et PETr sont comparables. Cependant, les moyennes des RCEQM relatives de HT-DFMTr sont systématiquement plus faibles que pour PETr. En outre, les taux de couverture moyens et les largeurs d'intervalle moyennes pour HT0-DFMTr sont meilleurs que ceux produits par HJ-DFMLin et HT-EBLUPLin. Par comparaison, pour le même ordre de biais relatif, la RCEQM relative de HT-EBLUPLin est plus faible que celle de HJ-DFMLin, et, bien que les deux estimateurs produisent des taux de couverture fort semblables, les intervalles de confiance générés par le HT-EBLUPLin ont tendance à avoir une largeur moyenne plus faible que ceux générés par HJ-DFMLin.

Les graphiques de la figure 1 représentent les mesures de performance selon la région obtenues pour ces quatre estimateurs pour l'ensemble A de simulations. Ces graphiques montrent que les valeurs du biais relatif et de la RCEQM relative générées par HT-DFMTr sont plus faibles que les valeurs correspondantes pour HJ-DFMLin et HT-EBLUPLin dans toutes les régions. Alors que les valeurs du biais relatif sont presque identiques, l'estimateur HT-DFMTr produit des valeurs plus faibles de la RCEQM relative que le PETr dans toutes les régions. En outre, le biais relatif et la RCEQM relative de HJ-DFMLin et HT-EBLUPLin augmentent à mesure que s'accroît la non-linéarité des données (c'est-à-dire quand nous passons de l'ensemble de paramètres 1 à l'ensemble de paramètres 6). Nous constatons également que HT-DFMTr produit dans toutes les régions de meilleurs taux de couverture que HT-EBLUPLin et HJ-DFMLin.

Tableau 4
Biais relatif moyen (BRM), moyenne des RCEQM relatives (MRCEQM) et taux de couverture moyen (TCM) pour les simulations fondées sur le plan de sondage en utilisant les données de l'AAGIS. Les erreurs-types de simulation du BRM et de la MRCEQM sont données entre parenthèses

Critère	Estimateur	Moyenne de 29 régions	Moyenne de 28 régions
BRM, %	HT-DFMTr	1,96 (0,20)	1,92 (0,11)
	HJ-DFMLin	-2,13 (0,15)	-2,21 (0,12)
	HT-EBLUPLin	2,98 (0,18)	3,36 (0,16)
	Pseudo-EBLUP	4,01 (0,22)	4,41 (0,20)
MRCEQM, %	JL	1,89 (0,19)	2,23 (0,17)
	HT-DFMTr	21,93 (4,47)	17,41 (1,18)
	HJ-DFMLin	20,15 (3,80)	16,91 (2,20)
	HT-EBLUPLin	19,87 (1,78)	19,30 (1,63)
TCM	Pseudo-EBLUP	22,42 (2,52)	21,95 (2,46)
	JL	20,97 (1,48)	20,48 (1,31)
	HT-DFMTr	0,89	0,92
	HJ-DFMLin	0,93	0,93
	HT-EBLUPLin	0,85	0,85

d'estimer les coûts agricoles annuels moyens (coûts de caissés totaux (CDT), mesurés en \$A) dans chaque région en utilisant la taille de l'exploitation agricole (hectares) comme variable auxiliaire. Nous nous sommes servis de la même spécification du modèle mixte que dans Chandra et Chambers (2005). Cette spécification comprend un terme d'interaction (zone par taille) dans les effets fixes et la spécification d'une pente aléatoire pour l'effet de domaine. Sous sa forme linéaire, le modèle n'est pas très bien ajusté aux données d'échantillon de l'AAGIS. L'adéquation du modèle est améliorée (quoique marginalement) quand on utilise une spécification linéaire à échelle logarithmique. Nos résultats sont résumés au tableau 4.

6.3 Discussion des résultats des simulations

Dans le tableau 2, les très grandes valeurs du biais relatif moyen de HJ-DFMTr sous la pondération calée sur un modèle fondée sur un modèle est l'aspect le plus frappant. Les deux estimateurs donnant les meilleurs résultats en ce qui concerne le biais relatif sont HT-DFMTr, qui est fondé sur les mêmes pondérations que HJ-DFMTr, et PETr. Un examen en vue de déterminer la raison de la performance médiocre de HJ-DFMTr a révélé que la sommation des poids calés sur un modèle fondés sur un modèle (17) à l'intérieur des petits domaines produisait des estimations extrêmement variables des tailles de population de petit domaine, ce qui sous-entend que ces poids ne peuvent pas être considérés comme « polyvalents » – ils fonctionnent bien quand ils sont utilisés avec des variables qui sont raisonnablement corrélées avec la variable qui définit le modèle sous valeurs prédites, mais peut donner de mauvais résultats avec d'autres variables moins bien corrélées (par exemple la variable indicatrice d'inclusion du petit domaine). Nous constatons en outre que ce problème ne se pose pas avec les poids EBLUP empiriques « classiques » pour le total (6), car HJ-DFMLin produit des résultats cohérents pour les six scénarios examinés dans l'ensemble

Tableau 2 Biens relatifs moyen (BRM), moyenne des RCEQM relatives (MRCEQMR), taux de couverture moyen (TCM) et largeur d'intervalle moyen (LIM) pour l'ensemble A de simulations fondées sur un modèle

Critère	Estimateur	Ensemble de paramètres					
		1	2	3	4	5	6
BRM, %	HJ-DEMT ^r	-82,68	-95,02	-98,08	-98,50	-98,29	-99,00
	HT-DEMT ^r	0,09	0,10	-0,14	-0,25	-0,03	0,04
	PET ^r	0,08	0,09	-0,18	-0,48	-0,05	0,01
	HJ-DFML ⁱⁿ	12,01	4,09	-1,35	-5,54	-6,60	-9,88
	HT-EBLUP ^{lin}	13,39	5,18	-0,67	-5,24	-6,41	-9,67
MRCEQMR	HJ-DEMT ^r	4,80	1,39	1,25	1,44	1,42	1,62
	HT-DEMT ^r	0,15	0,26	0,45	0,64	0,66	0,91
	PET ^r	0,30	0,41	0,58	0,80	0,81	1,09
	HJ-DFML ⁱⁿ	1,11	1,41	1,85	1,99	2,06	2,69
	HT-EBLUP ^{lin}	0,79	0,54	0,64	0,92	0,93	1,31
TCM	HJ-DEMT ^r	0,99	0,98	0,97	0,95	0,94	0,92
	HT-DEMT ^r	0,94	0,91	0,89	0,89	0,89	0,88
	HJ-DFML ⁱⁿ	0,87	0,85	0,85	0,88	0,88	0,87
	HT-EBLUP ^{lin}	0,85	0,85	0,86	0,87	0,88	0,87
LIM	HJ-DEMT ^r	1 592	22 688	140 452	52 × 10 ⁴	35 × 10 ⁵	44 × 10 ⁶
	HT-DEMT ^r	219	4 414	34 105	14 × 10 ⁴	11 × 10 ⁵	15 × 10 ⁶
	HJ-DFML ⁱⁿ	1 005	19 232	139 420	57 × 10 ⁴	41 × 10 ⁵	56 × 10 ⁶
	HT-EBLUP ^{lin}	382	7 099	57 039	26 × 10 ⁴	21 × 10 ⁵	32 × 10 ⁶

Tableau 3 Biens relatifs moyen (BRM), moyenne des RCEQM relatives (MRCEQMR), taux de couverture moyen (TCM) et largeur d'intervalle moyen (LIM) pour l'ensemble B de simulations fondées sur un modèle

Critère	Estimateur	Ensemble de paramètres					
		1	2	3	4	5	6
BRM, %	HJ-DEMT ^r	4,92	0,66	0,14	-1,50	-8,75	-8,85
	HJ-DFML ⁱⁿ	-0,21	0,04	0,12	0,16	0,17	-0,77
	HT-EBLUP ^{lin}	-0,19	0,04	0,13	0,17	0,17	-0,77
MRCEQMR	HJ-DEMT ^r	0,38	0,35	0,33	0,37	0,41	0,56
	HJ-DFML ⁱⁿ	0,56	0,36	0,34	0,53	1,20	1,20
	HT-EBLUP ^{lin}	0,38	0,30	0,29	0,36	0,56	0,56
TCM	HJ-DEMT ^r	0,94	0,92	0,92	0,91	0,87	0,87
	HT-DEMT ^r	0,91	0,92	0,92	0,92	0,90	0,90
	HJ-DFML ⁱⁿ	0,93	0,94	0,94	0,93	0,92	0,92
	HT-EBLUP ^{lin}	0,06	2,70	214	33 442	13 × 10 ⁶	10 × 10 ⁶
LIM	HJ-DEMT ^r	0,04	2,50	211	29 070	5 × 10 ⁶	5 × 10 ⁶
	HJ-DFML ⁱⁿ	0,06	2,70	214	38 660	13 × 10 ⁶	10 × 10 ⁶
	HT-EBLUP ^{lin}	0,05	2,60	214	33 442	10 × 10 ⁶	10 × 10 ⁶

6.2 L'étude par simulation fondée sur le plan de sondage

Pour cette étude, nous nous sommes servis de la même population et des mêmes échantillons que dans les études par simulation décrites dans Chandra et Chambers (2005) et dans Chambers et Tzavidis (2006), qui étaient fondées sur des données obtenues auprès d'un échantillon de 1 652 entreprises agricoles qui avaient participé à l'Australian Agricultural and Grazing Industries Survey (AAGIS). Une population réaliste de 81 982 exploitations agricoles a été définie par échantillonnage avec remise à partir de l'échantillon original de 1 652 exploitations agricoles avec probabilités proportionnelles aux poids de sondage, qui étaient tous

strictement supérieurs à 1. En tout, 1 000 échantillons indépendants, chacun de taille $n = 1 652$, ont été tirés de cette population fixe par échantillonnage aléatoire simple sans remise dans des strates définies par les 29 régions agricoles australiennes représentées dans l'échantillon de l'AAGIS. Ces régions sont les petits domaines d'intérêt. Les tailles des échantillons régionaux ont été fixées à la même valeur que dans l'échantillon original, variant de 6 à 117, ce qui permet d'évaluer la performance des diverses méthodes d'estimation sur une gamme de tailles d'échantillon de petit domaine réalistes. Notons que les fractions d'échantillon-nage dans ces strates variaient également de manière disproportionnée, allant de 0,70 % à 15,87 %. L'objectif était

6.1 L'étude par simulation fondée sur un modèle

Les simulations fondées sur un modèle sont utilisées fréquemment pour illustrer la sensibilité d'une méthode d'estimation à la variation des hypothèses au sujet de la structure de la population d'intérêt. Ici, nous avons fixé la taille de la population à $N = 15\,000$ et avons produit aléatoirement les tailles de population de petit domaine $N_i, i = 1, \dots, D = 30$ de sorte que $\sum_i N_i = N$. Nous avons utilisé une taille globale d'échantillon de $n = 600$ avec un ensemble de tailles d'échantillon de petit domaine tel que ces tailles étaient proportionnelles aux tailles de population de petit domaine correspondantes. Ces tailles de population et d'échantillon propres aux domaines ont été maintenues fixes dans toutes les simulations. Les tailles de population et d'échantillon sont présentées au tableau 1a.

Tableau 1a
Tailles de population (N_i) et d'échantillon (n_i) propres au domaine pour la simulation fondée sur un modèle

Domaine	N_i	n_i	Domaine	N_i	n_i	Domaine	N_i	n_i
1	525	21	11	502	20	21	506	27
2	538	22	12	524	21	22	506	28
3	510	20	13	509	19	23	513	25
4	468	19	14	484	18	24	536	26
5	526	20	15	487	17	25	495	27
6	484	19	16	459	16	26	506	28
7	516	21	17	542	18	27	495	29
8	458	19	18	498	19	28	463	30
9	529	21	19	512	20	29	497	460
10	518	21	20	500	20	30	497	460

Dans l'ensemble A de nos simulations fondées sur un modèle, les valeurs de population y_{ij} ont été produites en utilisant le modèle multiplicatif $y_{ij} = 5,0x_{ij}^{\beta}u_i e_{ij}^{\gamma}$ ($j = 1, \dots, N_i; i = 1, \dots, 30$), puis des échantillons aléatoires ont été tirés de chaque petit domaine. Ici, les valeurs de x_{ij} ont été tirées indépendamment de la loi log normale $\log(x_{ij}) \sim N(6, \sigma_x^2)$, avec les effets individuels et les effets de domaine tirés indépendamment en tant que $\log(e_{ij}^{\gamma}) \sim N(0, \sigma_e^2)$ et $\log(u_i^{\gamma}) \sim N(0, \sigma_u^2)$, respectivement. Le valeurs de population de x ont été générées à nouveau dans chaque simulation. En particulier, dans chaque simulation, nous avons d'abord généré les valeurs de x pour une population de taille N , puis attribué aléatoirement ces valeurs à divers domaines de tailles N_i . Les valeurs de σ_e^2 et σ_u^2 ont été choisies de manière que la corrélation intra-domaine dans la population varie entre 0,20 et 0,25. Le

tableau 1b montre les six ensembles distincts de valeurs des paramètres qui ont été utilisés dans l'ensemble A. Ces divers ensembles ont assuré que les populations simulées contiennent une grande gamme de variations. Pour chaque population générée et pour chaque domaine i , nous avons sélectionné un échantillon aléatoire simple (sans remise) de taille n_i , ce qui a donné une taille globale d'échantillon de $n = 600$. Les valeurs d'échantillon de y et les valeurs de population de x obtenues dans chaque simulation ont ensuite été utilisées pour estimer les moyennes de petit domaine. Autrement dit, en utilisant les données d'échantillon dans chaque cas, nous avons estimé les valeurs des paramètres en utilisant la fonction *lme* dans R (Bates et Pinheiro 1998), puis nous avons calculé les estimations des moyennes de petit domaine, ainsi que les intervalles de confiance à 95 % nominaux appropriés. Le processus de génération des données de population et d'échantillon, d'estimation des paramètres et de calcul des estimations sur petits domaines a été répété indépendamment 1 000 fois. Les résultats de cette partie de l'étude par simulation sont présentés au tableau 2.

Tableau 1b
Spécifications des populations pour l'ensemble A de simulations fondées sur un modèle

Ensemble de paramètres		
β	σ_u^2	σ_e^2
1	0,5	0,30
2	0,8	0,35
3	1,0	0,40
4	1,3	0,45
5	1,5	0,50
6	2,0	0,60
1	0,50	1,00
2	0,60	1,20
3	0,70	2,25
4	0,80	1,75
5	0,90	1,50
6	1,00	1,20

Dans l'ensemble B de simulations fondées sur un modèle, les données de population ont été générées en utilisant le modèle $y_{ij} = 5,0x_{ij}^{\beta}[\exp(\log^2(x_{ij}))]^{\gamma}u_i e_{ij}^{\gamma}$. Ici, les effets individuels e_{ij} et les effets de domaine u_i ont été tirés indépendamment en tant que $\log(e_{ij}^{\gamma}) \sim N(0, 1)$ et $\log(u_i^{\gamma}) \sim N(0, 0,25)$, respectivement, tandis que les valeurs des covariables x_{ij} ont été tirées en tant que $\log(x_{ij}) \sim N(3, 0,04)$. Cinq valeurs différentes du paramètre γ (-1,0, -0,5, 0,0, 0,5, 1,0) ont été examinées, ce qui a produit des données de population ayant divers degrés de courbure. Tous les autres aspects de ces simulations, y compris les estimateurs pris en considération, étaient les mêmes que pour l'ensemble A. Le tableau 3 donne les résultats pour cette composante de l'étude par simulation.

$$m_{\text{HJ-DFMTr}}^{ty} = \left\{ \sum_{j \in s_i} w_{\text{confme}}^{ij} \right\}^{-1} \sum_{j \in s_i} w_{\text{confme}}^{ij} y_{ij} \quad (18)$$

et

$$m_{\text{HT-DFMTr}}^{ty} = N_i^{-1} \sum_{j \in s_i} w_{\text{confme}}^{ij} y_{ij} \quad (19)$$

Nous pouvons aussi adopter une approche fondée sur la prédiction pour obtenir un prédicteur indirect de rechange de la moyenne de petit domaine sous le modèle transformé logarithmiquement (9). Notre approche étend celle de Karlberg (2000a). Dans ces conditions, en supposant que le modèle (9) est vérifié, nous prédisons chaque valeur de Y hors échantillon dans le petit domaine i , puis nous additionnons ces prédictions. Notons que nous devons corriger le biais après la rétrotransformation à l'échelle originale quand nous calculons ces valeurs prédites pour les valeurs de Y hors échantillon. Sous le modèle (9), le prédicteur empirique résultant de la moyenne m_{ty}^{ty} de Y dans le domaine i (désigné PETr) peut être défini comme étant

$$m_{\text{PETr}}^{ty} = N_i^{-1} \left\{ \sum_{j \in s_i} y_{ij} + \sum_{j \in r_i} \hat{y}_{ij} \right\}, \quad (20)$$

où \hat{y}_{ij} est donné par (15).

L'estimation de l'EQM de (18) et de (19) est effectuée de la façon habituelle pour les estimateurs DFM, c'est-à-dire suivant l'approche d'estimation de l'EQM décrite à la section 2. L'estimation de l'EQM de (20) n'est pas facile, parce que ce prédicteur est une fonction non linéaire des valeurs de Y . Nous ne poursuivons pas l'examen de cette question dans le présent article.

6. Une évaluation empirique

À la présente section, nous présentons les résultats empiriques d'une étude comparative des performances de cinq méthodes différentes d'estimations sur petits domaines. Ces méthodes comprennent les deux estimateurs directs fondés sur un modèle (DFM) « basé sur une transformation » (18) et (19), et s'appuient tous les deux sur les poids calculés sur un modèle fondés sur un modèle (17) et sont désignées par HJ-DFMTr et HT-DFMTr, respectivement ; le prédicteur (20) fondé sur une transformation logarithmique sous le modèle (9), désigné par PETr, l'estimateur DFM « classique » (7) basé sur le modèle linéaire mixte (3) et les poids empiriques EBLUP pour le total (6), que nous désignons par HJ-DFMLin afin de mettre l'accent sur le fait qu'il s'agit d'une moyenne pondérée de type Hájek fondée sur les poids calculés sous un modèle linéaire mixte, ainsi que l'estimateur EBLUP (8) calculé sous le même modèle

Nos résultats empiriques sont fondés sur deux types d'études par simulation. Le premier comprend l'utilisation d'une simulation fondée sur un modèle pour générer une population artificielle et des données d'échantillon. Autrement dit, à chaque simulation, nous avons d'abord généré des données de population sous le modèle, puis tiré un échantillon unique de cette population simulée par échantillonnage aléatoire simple stratifié sans remise en utilisant les petits domaines comme strates. Nous avons ensuite utilisé ces données pour comparer les propriétés des divers estimateurs. À la section 6.1, nous présentons les résultats de ces simulations fondées sur un modèle. Nous avons exécuté deux ensembles de ces simulations. Dans le premier (ensemble A), nous avons étudié la performance de ces estimateurs étant donné les données de population générées en utilisant le modèle linéaire mixte à échelle logarithmique (9). Dans le deuxième ensemble de simulations (ensemble B), nous avons examiné la robustesse de ces estimateurs à l'erreur de spécification de ce modèle. Le deuxième type d'études par simulation était fondé sur le plan de sondage. À la section 6.2, nous décrivons les simulations fondées sur le plan de sondage. Ici, nous avons évalué les estimateurs dans le contexte de l'échantillonnage répété à partir d'une population réelle en utilisant des méthodes d'échantillonnage réalistes. Autrement dit, nous avons d'abord utilisé des données d'enquête réelles pour simuler une population, puis nous avons échantillonné cette population fixe à plusieurs reprises conformément à un plan préspecifié. En particulier, nous avons utilisé un plan d'échantillonnage aléatoire stratifié dans lequel les strates correspondaient aux petits domaines d'intérêt et la répartition de l'échantillon entre les strates correspondait aux tailles d'échantillon de petit domaine dans les ensembles de données originaux.

Nous avons calculé quatre mesures de la performance des estimateurs en nous servant de diverses estimations produites dans les études par simulation. Ces mesures sont le biais relatif (BR) et la racine carrée de l'erreur quadratique moyenne (RCEQM) relative de ces estimations, ainsi que le taux de couverture et la largeur moyenne des intervalles de confiance à 95 % nominaux qui sont fondés sur ces taux. Aux tableaux 2 à 4, ces mesures sont présentées sous forme de moyennes sur les petits domaines d'intérêt.

nulle (McCulloch et Searle 2001, chapitre 2, pages 40-45), la covariance entre $\hat{\beta}$ et \hat{v}_{ijk}^s sera négligeable. Il s'ensuit que

$$tr[E\{z^{(2)}(\eta_j)(\hat{\eta}_j - \eta_j)(\hat{\eta}_j - \eta_j)\}] = tr[z^{(2)}(\eta_j)E\{(\hat{\eta}_j - \eta_j)(\hat{\eta}_j - \eta_j)\}]$$

$$\approx e^{\phi_{ij} + \frac{\psi_{ij}^2}{2}} \left[\mathbf{d}_{ij}' \left(\sum_{ss} \mathbf{d}_{ss}^s \hat{\mathbf{v}}_{ss}^{-1} \mathbf{d}_{ss}^s \right)^{-1} \mathbf{d}_{ij} + \frac{4}{1} \text{Var}(\hat{v}_{ij}^s) \right] = E(y_{ij} | x_{ij}, \mathbf{g}_{ij}) \left[\hat{a}_{ij} + \frac{4}{1} \text{Var}(\hat{v}_{ij}^s) \right]$$

où $\hat{a}_{ij} = \mathbf{d}_{ij}' V(\hat{\beta}) \mathbf{d}_{ij}$ et $V(\hat{\beta}) = (\sum_{ls} \mathbf{d}_{ls}' \hat{\mathbf{V}}_{ls}^{-1} \mathbf{d}_{ls})^{-1}$ est l'estimateur habituel de $\text{Var}(\hat{\beta})$. Nos valeurs prédites sont par conséquent définies par l'estimateur corrigé du biais de deuxième ordre de $E(y_{ij} | x_{ij}, \mathbf{g}_{ij})$,

$$\hat{y}_{ij} = h(\mathbf{d}_{ij}; \eta_j) = \hat{k}_{ij}^{-1} e^{\phi_{ij} + \psi_{ij}^2/2} \quad (15)$$

où

$$\hat{k}_{ij} = 1 + \frac{1}{2} \left\{ \hat{a}_{ij} + \frac{4}{1} V(\hat{v}_{ij}^s) \right\}$$

et $V(\hat{v}_{ij}^s)$ est la variance asymptotique estimée de \hat{v}_{ij}^s . Sous estimation du MV et du MVR des composantes de la variance de (9), cette variance asymptotique estimée s'obtient à partir de l'inverse de la matrice d'information pertinente. Notons que la correction du biais de Karlberg (2000a) est un cas particulier de (15).

Afin d'utiliser (14) pour définir les poids de sondage calés sur un modèle fondé sur un modèle, nous avons également besoin des estimations des moments de deuxième ordre des valeurs de population de X sachant les valeurs prédites. Les moments conditionnels ω_{ijk} représentent une approximation de premier ordre de ces moments. En particulier, si les effets aléatoires sont normaux,

$$\omega_{ijk} = e^{(\phi_{ij} + \phi_{ijk}) + (\psi_{ij} + \psi_{ijk})/2} (e^{\psi_{ijk}} - 1) \quad (16)$$

Nous obtenons notre estimation $\hat{\omega}_{ijk}$ de ω_{ijk} en substituant $\hat{\phi}_{ij}$ et $\hat{\psi}_{ijk}$ à ϕ_{ij} et ψ_{ijk} dans (16). Les poids calés sur un modèle fondés sur un modèle empiriques (14) correspondants au modèle sous valeurs prédites défini par (15) et (16) sont

$$\mathbf{w}_{confine}^i = (w_{confine}^{ij}; j \in s_i^i; i = 1, \dots, D)$$

$$= \mathbf{1}^s + \mathbf{H}_i^s (\mathbf{J}_i^U \mathbf{1}^U - \mathbf{J}_i^s \mathbf{1}^s)$$

$$+ (\mathbf{1}^s - \mathbf{H}_i^s \mathbf{J}_i^s) \hat{\Omega}_{ss}^{-1} \hat{\Omega}_{ss}^s \mathbf{1}^s, \quad (17)$$

$$[c_i, \mathbf{J}^U] = [\mathbf{1}^U, \hat{\mathbf{y}}^U], \text{ donc}$$

$$\hat{\omega}_{ijk} = e^{\phi_{ij} + \phi_{ijk} + \hat{\sigma}_{ij}^2 + \hat{\sigma}_{ijk}^2} [e^{\hat{\sigma}_{ij}^2} \{1 + I(j = k)(e^{\hat{\sigma}_{ij}^2} - 1)\} - 1].$$

Jusqu'à présent, notre exposé s'est appuyé sur l'hypothèse de normalité des effets aléatoires sur l'échelle logarithmique. Toutefois, il n'existe aucune bonne raison (autre la commodité) de supposer que, dans le cas de données asymétriques, ces effets aléatoires de domaine doivent être normaux. Une autre option, étant donné un effet de domaine scalaire dans (9), consiste à supposer que les effets aléatoires dans ce modèle sont tirés de la famille de lois *gamma*.

Partant des propriétés de cette loi et en utilisant les développements des fonctions binomiales et exponentielles (en ignorant les termes d'ordre élevé), nous pouvons montrer que $E(y_{ij} | x_{ij}, \mathbf{g}_{ij}) \approx e^{\phi_{ij} + \psi_{ij}^2/2} = z(\eta_j)$, comme dans le cas normal. Cela indique qu'un estimateur DFM basé sur les poids calés sur un modèle fondés sur un modèle (17) devrait être robuste en ce qui concerne la distribution des effets aléatoires dans (9).

Enfin, nous considérons la définition de l'estimateur DFM proprement dit. Comme nous l'avons mentionné à la section 2, cet estimateur est simplement la moyenne pondérée des valeurs d'échantillon de Y dans un domaine. Cependant, l'utilisation de cette moyenne pondérée présuppose que les poids sont raisonnablement proches d'être calés localement sur N_i^i , autrement dit que, si nous effectuons la sommation sur les unités échantillonnées dans le domaine i , nous obtenons une valeur qui ne diffère pas trop de la taille de population réelle du petit domaine N_i^i . Cette propriété est habituellement vérifiée si les poids sont les poids EBLUP pour le total (6) défini par un modèle linéaire mixte pour Y . Elle ne l'est pas nécessairement pour les poids calés sur un modèle fondés sur un modèle (17). Par conséquent, nous considérons deux spécifications pour l'estimateur DFM sachant ces poids. La première, que nous appelons « spécification de Hájek », est simplement la moyenne pondérée (7), les poids étant définis par (17). La deuxième, que nous appelons « spécification d'Horvitz-Thompson », remplace le dénominateur de (7) par la valeur réelle de N_i^i . Autrement dit, les deux types d'estimateurs DFM sous la pondération calée sur un modèle fondée sur un modèle que nous considérons sont

$$\mathbf{J}_i^U \mathbf{1}^U - \mathbf{J}_i^s \mathbf{1}^s = \left(\sum_{i \in I} \hat{y}_{ij}^s \right) \cdot \begin{pmatrix} N - n \\ \vdots \end{pmatrix}$$

et $\mathbf{H}_{mc}^s = (\mathbf{J}_i^s \hat{\Omega}_{ss}^{-1} \mathbf{J}_i^s)^{-1} \mathbf{J}_i^s \hat{\Omega}_{ss}^{-1}$. En outre $\hat{\Omega}_{ss}^s = \text{diag}\{\hat{\Omega}_{ss}^{ls}; l = 1, \dots, D\}$ et $\hat{\Omega}_{ss}^{ls} = \text{diag}\{\hat{\Omega}_{ss}^{lstr}; l = 1, \dots, D\}$, où $\hat{\Omega}_{ss}^{lstr}$ et $\hat{\Omega}_{ss}^{ls}$ sont définis par la décomposition échantillon/hors échantillon de $\hat{\Omega}_{ss}^s$. Par exemple, quand (9) correspond à une spécification d'ordonnées à l'origine aléatoires, $\hat{v}_{ij}^s = \hat{\sigma}_{ij}^2 + \hat{\sigma}_{ij}^2 I(j = k)$ et les composantes de $\hat{\Omega}_{ss}^s$ sont donc

de population y_j et les valeurs prédites $\hat{y}_j = h(\mathbf{x}_j; \hat{\eta})$ semble être un point de départ raisonnable. Nous rem-plaçons par conséquent le modèle non linéaire (11) par le modèle linéaire

$$E(y_j | \hat{y}_j) = \alpha_0 + \alpha_1 \hat{y}_j$$

et

$$\text{Cov}(y_j, y_k | \hat{y}_j, \hat{y}_k) = \omega_{jk} \quad (13)$$

Nous désignons (13) comme étant le modèle « sous valeurs prédites » correspondant à (11). Soit \mathbf{J}^U la « matrice de plan » de population sous le modèle (13), c'est-à-dire $\mathbf{J}^U = [\mathbf{1}^U \ \hat{\mathbf{Y}}^U]$, où $\mathbf{1}^U$ désigne le vecteur unitaire de taille N et $\hat{\mathbf{Y}}^U = (\hat{y}_j; j = 1, \dots, N)$, et posons que $\Omega^U = [\omega_{jk}; j = 1, \dots, N; k = 1, \dots, N]$. Nous pouvons maintenant partitionner \mathbf{J}^U et Ω^U en fonction des unités échantillonées (s) et non échantillonées (r) comme il suit

$$\mathbf{J}^U = \begin{bmatrix} \mathbf{J}^r \\ \mathbf{J}^s \end{bmatrix}$$

$$\Omega^U = \begin{bmatrix} \Omega^{rs} & \Omega^{rs} \\ \Omega^{rs} & \Omega^{rr} \end{bmatrix}$$

et donc écrire les poids qui définissent l'estimateur BLUP de t^U sous (13). Il s'agit des poids calés sur un modèle fondé sur un modèle (cmln)

$$\mathbf{w}^{cmln} = (\mathbf{w}^{cmln}_j; j \in s)$$

$$= \mathbf{1}^s + \mathbf{H}^{cm}(\mathbf{J}^U \mathbf{1}^U - \mathbf{J}^s \mathbf{1}^s) + (\mathbf{1}^s - \mathbf{H}^{cm} \mathbf{J}^s) \Omega^{rs} \Omega^{sr} \mathbf{1}^r \quad (14)$$

où $\mathbf{H}^{cm} = (\mathbf{J}^s \Omega^{ss} \mathbf{J}^s)^{-1} \mathbf{J}^s \Omega^{ss}$. De toute évidence, ces poids sont calés sur un modèle puisque $\sum_{j \in s} \mathbf{w}^{cmln}_j = N$ et $\sum_{j \in s} \mathbf{w}^{cmln}_j \hat{y}_j = \sum_{j \in U} \hat{y}_j$. Cependant, contrairement aux poids EBLUP du modèle linéaire (2), ils ne sont pas calés sur \mathbf{X} . En pratique, les composantes de Ω^U sont inconnues et doivent être estimées. Quand ces estimations sont introduites par substitution dans (14), nous obtenons la version empirique \mathbf{w}^{cmlne} de ces poids calés sur un modèle.

5. Pondération calée sur un modèle pour l'estimation sur petits domaines

Nous utilisons maintenant le calage sur un modèle fondé sur le modèle linéaire mixte à échelle logarithmique (9) pour obtenir les poids de sondage à utiliser dans l'estimateur DFM (7). L'exposé de la section précédente montre que, pour cela, nous devons d'abord spécifier un modèle sous valeurs prédites (13) pour Y basé sur (9), c'est-à-dire que nous devons calculer les valeurs prédites appropriées \hat{y}_{ij}

de sorte que la correction habituelle du biais s'appuyant sur le fait que la loi conditionnelle de y_{ij} est lognormale est inadéquate. Soit $\hat{\eta}_{ij} = (\beta, \hat{v}_{ij})'$ une estimation de $\eta_{ij} = (\beta, v_{ij})'$ telle que $E(\hat{\eta}_{ij}) = \eta_{ij}$. En utilisant une approximation par développement en série de Taylor de deuxième ordre, nous pouvons écrire

$$z(\hat{\eta}_{ij}) \approx z(\eta_{ij}) + (\hat{\eta}_{ij} - \eta_{ij})' z^{(1)}(\eta_{ij})$$

$$+ \frac{1}{2} (\hat{\eta}_{ij} - \eta_{ij})' z^{(2)}(\eta_{ij}) (\hat{\eta}_{ij} - \eta_{ij})$$

et donc

$$E\{z(\hat{\eta}_{ij})\} \approx z(\eta_{ij})$$

$$+ \frac{1}{2} tr[E\{z^{(2)}(\eta_{ij})(\hat{\eta}_{ij} - \eta_{ij})(\hat{\eta}_{ij} - \eta_{ij})'\}]$$

Ici

$$z^{(1)}(\eta_{ij}) = \left(\mathbf{d}_{ij} e^{\phi_{ij} + v_{ij}/2} \frac{1}{2} e^{\phi_{ij} + v_{ij}/2} \right)$$

et

$$z^{(2)}(\eta_{ij}) = \begin{pmatrix} \mathbf{d}_{ij} \mathbf{d}_{ij}' e^{\phi_{ij} + v_{ij}/2} & \frac{1}{2} \mathbf{d}_{ij} e^{\phi_{ij} + v_{ij}/2} \\ \frac{1}{2} \mathbf{d}_{ij} e^{\phi_{ij} + v_{ij}/2} & \frac{1}{4} e^{\phi_{ij} + v_{ij}/2} \end{pmatrix}$$

sont le vecteur et la matrice, respectivement, contenant les dérivées de premier et de deuxième ordre de $z(\eta_{ij})$ par rapport à η_{ij} . Puisque la covariance asymptotique entre les estimateurs MV (ou MVR) des composantes fixes et des composantes de variance d'un modèle linéaire mixte est

caractériser la façon dont la régression de $\log(Y)$ sur $\log(X)$ varie d'un petit domaine à l'autre. Autrement dit, pour $i = 1, \dots, D$; $j = 1, \dots, N_i$ nous avons

$$l_{ij} = \log(y_{ij}) = \beta_0 + \beta_1 \log(x_{ij}) + \mathbf{g}_{ij}' \mathbf{u}_i + e_{ij} \quad (9)$$

où y_{ij} et x_{ij} sont les valeurs de Y et X , respectivement, pour l'unité de population j dans le petit domaine i , \mathbf{g}_{ij} désigne une covariable « contextuelle » de dimension q , \mathbf{u}_i désigne un effet aléatoire pour le domaine i , également de dimension q , et e_{ij} est un effet aléatoire individuel scalaire. Comme d'habitude pour ce genre de modèle, nous supposons que tous les effets aléatoires suivent une loi normale et sont mutuellement non corrélés, que leurs valeurs espérées sont nulles et que $\text{Var}(\mathbf{u}_i) = \Sigma_u$ et $\text{Var}(e_{ij}) = \sigma_e^2$. Ici, Σ_u est la matrice de dimensions $q \times q$ de covariance des effets aléatoires. Notons que $\text{Var}(l_{ij} | x_{ij}) = \mathbf{g}_{ij}' \Sigma_u \mathbf{g}_{ij} + \sigma_e^2$ et $\text{Cov}(l_{ij}, l_{ik} | x_{ij}, x_{ik}, \mathbf{g}_{ij}, \mathbf{g}_{ik}) = \mathbf{g}_{ij}' \Sigma_u \mathbf{g}_{ik}$ sous (9).

Sachant les valeurs d'échantillon de y_{ij} , x_{ij} et \mathbf{g}_{ij} , nous pouvons appliquer des méthodes d'estimation classiques (par exemple, maximum de vraisemblance (MV) ou maximum de vraisemblance restreint (MVR), voir Harville 1977) pour estimer les paramètres de (9). Soit Σ_u et σ_e^2 les estimations résultantes des composantes de la variance de ce modèle linéaire mixte. L'estimation de $\beta = (\beta_0, \beta_1)'$ est alors

$$\hat{\beta} = \left(\sum_i \mathbf{d}_i' \mathbf{V}_i^{-1} \mathbf{d}_i \right)^{-1} \left(\sum_i \mathbf{d}_i' \mathbf{V}_i^{-1} \mathbf{y}_i \right) \quad (10)$$

où $\mathbf{V}_i = \mathbf{V}_i^{ls}$, $\mathbf{d}_i = \mathbf{d}_i^{ls}$ et $\mathbf{V}_i^{ls} = \mathbf{V}_i^{ls} + \sigma_e^2 \mathbf{I}_i$ sont les composantes d'échantillon de $\mathbf{V}_i = [\mathbf{v}_{ijk}] = \mathbf{g}_i' \Sigma_u \mathbf{g}_i + \sigma_e^2 \mathbf{I}_i$, $\mathbf{d}_i = [\mathbf{d}_{ijk}] = [\mathbf{1}_i \log(x_i)]$ et $\mathbf{I}_i = (\mathbf{I}_{ij})$; $j = 1, \dots, N_i$, respectivement. Ici, \mathbf{g}_i est la matrice de dimensions $N_i \times q$ définie par les covariables \mathbf{g}_{ij} dans le domaine i , \mathbf{I}_i est la matrice identité d'ordre N_i , $\mathbf{1}_i$ désigne un vecteur de valeurs 1 de dimension N_i et $\log(x_i)$ désigne le vecteur de N_i valeurs de $\log(X)$ dans le domaine i .

Notons que, quand les composantes de la variance Σ_u et σ_e^2 sont connues, (10) est l'estimateur BLUE pour β . Par conséquent, $E(\hat{\beta}) \approx \beta$ et $\text{Var}(\hat{\beta}) \approx (\sum_i \mathbf{d}_i' \mathbf{V}_i^{ls} \mathbf{d}_i)^{-1}$. Posons que $\hat{\phi}_i = (\hat{\phi}_{ij}) = \mathbf{d}_i' \hat{\beta}$. Alors $E(\hat{\phi}_i) \approx \mathbf{d}_i' \beta$ et $\text{Var}(\hat{\phi}_i) = \mathbf{A}_i = [\mathbf{a}_{ijk}] \approx \mathbf{d}_i' (\sum_i \mathbf{d}_i' \mathbf{V}_i^{ls} \mathbf{d}_i)^{-1} \mathbf{d}_i$, où $\mathbf{a}_{ijk} = \mathbf{d}_{ij}' \text{Var}(\hat{\beta}) \mathbf{d}_{ik} \rightarrow 0$ quand $n \rightarrow \infty$.

Notre but est d'utiliser le modèle linéaire mixte à échelle logarithmique (9) pour estimer les moyennes de petit domaine m_{ij} . En particulier, nous utilisons le calage sur un modèle (Wu et Sitter 2001) basé sur ce modèle pour calculer les poids de sondage à utiliser dans l'estimateur DFM (7) de cette quantité.

4. Pondération calée sur un modèle

Le calage sur un modèle a été introduit par Wu et Sitter (2001) en tant que méthode de pondération par calage

Supposons que le modèle de population sous-jacent est non linéaire, la relation entre Y et X dans la population étant de la forme

$$E(y_j | \mathbf{x}_j) = h(\mathbf{x}_j; \eta) \text{ et } \text{Var}(y_j | \mathbf{x}_j) = \sigma_j^2 \quad (11)$$

Ici, $j = 1, \dots, N$, η (habituellement évalué vectoriellement) et σ_j^2 sont les paramètres inconnus du modèle, et la fonction de la moyenne $h(\mathbf{x}_j; \eta)$ est une fonction connue de \mathbf{x}_j et η . Nous supposons aussi que les unités de population sont mutuellement non corrélées, sachant leurs valeurs respectives de X . Notons que (11) est une expression assez générale qui englobe les modèles linéaire, non linéaire et linéaire généralisé comme cas particuliers. Dans cette situation, Wu et Sitter (2001) définissent l'estimateur calé sur un modèle (cm) du total de population t_{ycm}^s comme étant $t_{ycm}^s = \sum_{j \in s} w_{cm}^j y_j$, où le vecteur de poids $w_{cm}^s = (w_{cm}^j)$ est choisi de manière à minimiser une mesure appropriée de la distance entre $w_{cm}^s = (\pi_j^{-1})$, sous les contraintes de

$$\sum_{j \in s} w_{cm}^j = N$$

et

$$\sum_{j \in s} w_{cm}^j h(\mathbf{x}_j; \hat{\eta}^\pi) = \sum_{j \in U} h(\mathbf{x}_j; \hat{\eta}^\pi) \quad (12)$$

où $\hat{\eta}^\pi$ est un estimateur convergent sous le plan de η . Notons que, contrairement au calage classique, les contraintes (12) exigent que nous connaissions les valeurs de population individuelles de X . L'idée essentielle sur laquelle s'appuie cette approche est que, à condition que (11) soit ajusté raisonnablement, y_j est (du moins approximativement) une fonction linéaire de sa valeur prédite $h(\mathbf{x}_j; \hat{\eta}^\pi)$ sous ce modèle et que nous pouvons donc exécuter l'estimation linéaire en utilisant ces valeurs prédites comme information auxiliaire.

Une perspective fondée sur un modèle du calage sur un modèle peut être élaborée comme il suit. Soit $\hat{\eta}$ un estimateur « efficace sous le modèle » de η dans (11), par exemple son estimateur par le maximum de vraisemblance (MV), avec les valeurs prédites $h(\mathbf{x}_j; \hat{\eta})$. En général, ces valeurs prédites ne sont pas sans biais. Elles sont également corrélées. Toutefois, il persiste une relation systématique entre les valeurs réelles de Y et les valeurs prédites correspondantes que nous pouvons approximer. Bien que rien ne nous empêche d'examiner des approximations plus complexes, un modèle linéaire pour la relation entre les valeurs

$$\mathbf{V}^{sr} = \text{diag}\{\mathbf{V}^{lsr}; i = 1, \dots, D\}$$

$$= \text{diag}\{\mathbf{g}_{is}^s \Sigma^n \mathbf{g}_{is}^{sr}; i = 1, \dots, D\}. \quad (5)$$

Ici, \mathbf{g}_{is} et \mathbf{g}_{is}^{sr} désignent la restriction de \mathbf{g}_i^l aux unités échantillonnées et non échantillonnées dans le domaine i , respectivement. Sachant les valeurs estimées $\hat{\theta} = (\hat{\Sigma}^n, \hat{\sigma}^2)$ des composantes de la variance, nous pouvons les introduire par substitution dans (4) et (5) pour obtenir les estimations $\hat{\mathbf{V}}^{ss}$ et $\hat{\mathbf{V}}^{sr}$ de \mathbf{V}^{ss} et \mathbf{V}^{sr} , respectivement, et par conséquent calculer les poids BLUP « empiriques », ou poids EBLUP, pour le total de population de Y sous la forme

$$\mathbf{w}_{EBLUP}^s = (\mathbf{w}_{EBLUP}^{ij})_{j \in s_i; i = 1, \dots, D}$$

$$= \mathbf{1}_i + \mathbf{H}_i'(\mathbf{t}_i - \mathbf{t}_i^*)$$

$$+ (\mathbf{I}_s - \mathbf{H}_i' \mathbf{x}_i^s) \hat{\mathbf{V}}^{ss-1} \hat{\mathbf{V}}^{sr} \mathbf{1}_r \quad (6)$$

où $\mathbf{H}_s = (\mathbf{x}_i^s \hat{\mathbf{V}}^{ss-1} \mathbf{x}_i^s)^{-1} \mathbf{x}_i^s \hat{\mathbf{V}}^{ss-1}$. Notons que nous utilisons maintenant le double indice inférieur ij afin de faire la distinction entre les unités de population dans les différents domaines.

L'estimateur DFM de la moyenne m_{ij}^l de Y dans le domaine i (Chandra et Chambers 2005, 2009) fondé sur les poids EBLUP pour le total (6) est simplement la moyenne pondérée correspondante des valeurs d'échantillon de Y dans le domaine i ,

$$m_{ij}^{HD-DFMLin} = \left\{ \sum_{j \in s_i} \mathbf{w}_{EBLUP}^{ij} \right\}^{-1} \sum_{j \in s_i} \mathbf{w}_{EBLUP}^{ij} y_{ij}^l. \quad (7)$$

Notons que (7) n'est pas l'EBLUP pour m_{ij}^l sous (3). Celui-ci est (voir Rao 2003, section 6.2.3)

$$m_{ij}^{HT-EBLUPLin}$$

$$= E\{m_{ij}^l | \mathbf{y}^{ls}, \mathbf{x}^{ls}, \mathbf{x}^{lr}\}$$

$$= N_i^{-1} \left[\sum_{j \in s_i} y_j^l + \mathbf{1}_{lr}' \left\{ \mathbf{x}_{lr}^l \hat{\beta} + \hat{\mathbf{V}}^{lr} \hat{\mathbf{V}}^{ls-1} (\mathbf{y}^{ls} - \mathbf{x}_{ls}^l \hat{\beta}) \right\} \right]$$

$$= N_i^{-1} \left[n_{i, \bar{y}}^l + (N_i - n_i^l) \left\{ \mathbf{x}_{lr}^l \hat{\beta} + \hat{\mathbf{g}}_{lr}^l \hat{\Sigma}^n \mathbf{g}_{lr}^{ls} (\hat{\mathbf{g}}_{ls}^s \hat{\Sigma}^n \mathbf{g}_{lr}^{ls} + \hat{\sigma}^2 \mathbf{1}_{ls}^l)^{-1} (\mathbf{y}^{ls} - \mathbf{x}_{ls}^l \hat{\beta}) \right\} \right]. \quad (8)$$

Ici, E désigne l'opérateur d'espérance sous (3) avec les paramètres inconnus remplacés par les estimations, \mathbf{x}_{ls}^l et \mathbf{x}_{lr}^l sont les matrices d'échantillon et hors échantillon de \mathbf{X} dans le domaine i , \mathbf{y}^{ls} est le vecteur des valeurs d'échantillon de Y dans le même domaine, $\hat{\beta}$ est le prédicteur BLUP « empirique » de β , $\hat{\mathbf{V}}^{lr}$ est la transposée de la valeur estimée de \mathbf{V}^{lr} avec $\hat{\mathbf{V}}^{lsr}$ l'estimation correspondante de \mathbf{V}^{lsr} , voir (4) et (5), et $\mathbf{1}_{lr}^l$ est un vecteur de valeurs 1 de longueur $N_i - n_i^l$. Notons que la dernière expression dans le deuxième membre de (8) découle directement de la substitution de (4) et (5), avec $\hat{\mathbf{x}}_{lr}^l$ et $\hat{\mathbf{g}}_{lr}^l$ désignant les vecteurs colonne d'ordre p et q définis en calculant la

moyenne des colonnes de \mathbf{x}_{lr}^l et \mathbf{g}_{lr}^l , respectivement. Comme le prédicteur EBLUP (8), l'estimateur (7) est une fonction pondérée de toutes les valeurs d'échantillon. Notons que sous la spécification de l'ordonnée à l'origine aléatoire de (3), (8) se réduit à l'expression (7.2.39) dans Rao (2003, section 7.2).

L'estimation de l'erreur quadratique moyenne (EQM) pour (8) est habituellement effectuée en appliquant la théorie décrite dans Prasad et Rao (1990). Bien que cet estimateur de l'EQM soit assez compliqué, il donne de bons résultats sous (3). Toutefois, si le modèle (3) n'est pas vérifié, les résultats peuvent être erronés. Il ne convient pas non plus comme estimateur de l'EQM de (8), sous échantillonnage répété, comme l'a fait remarquer Longford (2007). En revanche, l'estimation de l'EQM de (7) est assez simple. En effet, si l'on traite les poids qui définissent cet estimateur comme étant fixes, il s'agit d'un estimateur linéaire d'une moyenne de domaine et, par conséquent, sa variance de prédiction V_i sous (1) peut être estimée au moyen de méthodes bien connues (voir Royall et Cumberland 1978). Puisqu'en général les poids EBLUP pour le total (6) ne sont pas « calés localement » (c'est-à-dire qu'ils ne reproduisent pas la moyenne $\bar{\mathbf{x}}_i$ du domaine i de \mathbf{X}), (7) possède un biais B_i sous (1). Une simple estimation de ce biais par introduction des valeurs est la différence entre (7) et $\bar{\mathbf{x}}_i' \hat{\beta}$. L'estimateur final de l'EQM utilisé avec (7) est par conséquent défini en additionnant l'estimation de V_i et le carré de cette estimation de B_i . On a démontré empiriquement que cette méthode d'estimation de l'EQM possède de bonnes propriétés fondées sur un modèle ainsi que sous échantillonnage répété. Voir Chandra et Chambers (2005, 2009), Chambers et Tzavidis (2006), Chandra, Salvari et Chambers (2007), ainsi que Tzavidis, Salvari, Pratesi et Chambers (2008).

3. Estimation sur petits domaines sous transformation

À la présente section, nous étendons l'approche DFM à l'estimation sur petits domaines quand les relations de régression sous-jacentes ne sont pas linéaires. Ce faisant, nous nous concentrerons sur le cas important où les valeurs de population de Y suivent un modèle non linéaire sur leur échelle originale (brute), mais où leurs logarithmes peuvent être modélisés linéairement. L'extension à d'autres modèles de linéarisation est simple.

Sans perte de généralité, nous supposons que Y et X sont toutes deux des grandeurs scalaires strictement positives, dont la distribution marginale de population est asymétrique et pour lesquelles existent des preuves manifestes que leur relation est non linéaire, comme cela est le cas pour de nombreuses applications aux enquêtes auprès des entreprises. En outre, un modèle linéaire mixte convient pour

EBLUP ainsi qu'à l'estimateur DFM « habituel » défini en ajustant un modèle linéaire mixte aux données, ainsi qu'un prédicteur empirique indirect basé sur le même modèle linéaire mixte sous échelle transformée. À la section 7, nous concluons l'article par une discussion des questions en suspens.

2. Estimation directe fondée sur un modèle pour petits domaines

Il convient de souligner que l'approche adoptée dans le présent article est fondée sur un modèle. Par conséquent, tous les moments sont évalués par rapport à un modèle pour les données de population. En outre, nous supposons que toutes les données d'échantillon ont été obtenues par une méthode d'échantillonnage non informatif, par exemple l'échantillonnage probabiliste avec probabilités d'inclusion définies par des covariables connues du modèle.

Pour commencer, nous établissons la notation. Soit U une population de taille N et soit \mathbf{y}^U , le vecteur de dimension N des valeurs de population d'une caractéristique Y d'intérêt. Supposons que notre principal objectif soit l'estimation du total $t_{Yj} = \sum \mathbf{y}^U_j$ de ces valeurs de population (ou de leur moyenne $m_{Yj} = N^{-1} \sum \mathbf{y}^U_j$). Soit \mathbf{X} un vecteur de dimension p de variables auxiliaires qui sont reliées, d'une certaine façon, à Y et soit \mathbf{x}^U , la matrice de dimensions $N \times p$ correspondante des valeurs de population de ces variables. Nous supposons que les valeurs d'échantillon individuelles de \mathbf{X} sont connues. Les valeurs hors échantillon de \mathbf{X} ne sont pas nécessairement connues individuellement, mais sont supposées connues à un certain niveau d'aggrégation. Au minimum, nous connaissons le vecteur des totaux de population \mathbf{t}^U des colonnes de \mathbf{X} .

Supposons qu'il est raisonnable d'émettre l'hypothèse que la régression de Y sur \mathbf{X} au sein de la population est linéaire, c'est-à-dire que

$$E(\mathbf{y}^U | \mathbf{x}^U) = \mathbf{x}^U \beta \text{ et } \text{Var}(\mathbf{y}^U | \mathbf{x}^U) = \mathbf{V}^U \quad (1)$$

où \mathbf{V}^U est connu jusqu'à une constante multiplicative. Étant donné un échantillon s de taille n pour cette population, nous pouvons partitionner

$$\mathbf{V}^U = \begin{bmatrix} \mathbf{V}^{us} & \mathbf{V}^{us'} \\ \mathbf{V}^{us} & \mathbf{V}^{ss'} \end{bmatrix}$$

en leurs composantes dans l'échantillon et hors échantillon. Ici, $r = U - s$ désigne les unités de population qui ne sont

pas incluses dans l'échantillon. Le vecteur des poids qui définissent le meilleur prédicteur linéaire sans biais (BLUP pour *Best Linear Unbiased Predictor*) de t_{Yj} est alors (voir Royall 1976 ; Valliant, Dorfman et Royall 2000, section 2.4)

$$\mathbf{w}_{\text{BLUP}}^s = (\mathbf{w}_{\text{BLUP}}^j : j \in s)$$

$$= \mathbf{I}_s + \mathbf{H}^s((\mathbf{t}^r - \mathbf{t}^{ss}) + (\mathbf{I}_r - \mathbf{H}^r(\mathbf{x}^r)) \mathbf{V}^{rs} \mathbf{V}^{rr} \mathbf{I}_r) \quad (2)$$

où $\mathbf{H}^s = (\mathbf{x}_s^s \mathbf{V}^{ss} \mathbf{x}_s^s)^{-1} \mathbf{x}_s^s \mathbf{V}^{ss} \mathbf{I}_s$ est la matrice identité d'ordre n , \mathbf{t}^{ss} est le vecteur des totaux d'échantillon de \mathbf{X} et \mathbf{I}_s (\mathbf{I}_r) désigne un vecteur de valeurs 1 de taille n ($N - n$).

Nous supposons maintenant que la population cible U de taille N peut être partitionnée en D petits domaines non chevauchants, chacun de taille N_i , $i = 1, \dots, D$, tels que $N = \sum_{i=1}^D N_i$. Sachant qu'un échantillon s de n unités est tiré de cette population, nous supposons qu'un sous-échantillon s_i de n_i unités est tiré du domaine i , avec $n = \sum_{i=1}^D n_i$. Il convient de noter que nous supposons que tous les petits domaines sont échantillonnés et qu'il existe au moins une unité échantillonnée dans chaque petit domaine d'intérêt. Comme nous l'avons mentionné à la section 1, des modèles linéaires mixtes sont souvent utilisés pour l'estimation sur petits domaines. Ces modèles peuvent s'écrire sous la

$$\mathbf{y}^U = \mathbf{x}^U \beta + \mathbf{g}^U \mathbf{u} + \mathbf{e}^U \quad (3)$$

où \mathbf{u} est un vecteur aléatoire d'effets dits de domaine, \mathbf{e}^U est un vecteur de population de dimension N d'effets individuels aléatoires et \mathbf{g}^U est une matrice connue. En général, les effets de domaine sont évalués vectoriellement, de sorte que $\mathbf{u} = (\mathbf{u}_1^T \dots \mathbf{u}_q^T)^T$ et $\mathbf{g}^U = \text{diag}\{\mathbf{g}_i : i = 1, \dots, D\}$, où \mathbf{g}_i est de dimensions $N_i \times q$. Nous supposons que les effets particuliers au domaine $\{\mathbf{u}_i : i = 1, \dots, D\}$ sont des réalisations indépendantes et identiquement distribuées d'un vecteur aléatoire de dimension q dont la moyenne est nulle et la matrice de covariance est Σ^u . De même, nous supposons que les effets individuels scalaires constituant \mathbf{e}^U sont des réalisations indépendantes et identiquement distribuées d'une variable aléatoire de moyenne nulle et de variance σ_e^2 , les effets de domaine et les effets individuels étant mutuellement indépendants. Les paramètres $\theta = (\Sigma^u, \sigma_e^2)$ sont habituellement appelés les composantes de la variance de (3).

Étant donné les valeurs des composantes de la variance, il est facile de voir que (3) est simplement un cas particulier du modèle linéaire général (1) qui sous-tend les poids BLUP

$$\mathbf{V}^{ss} = \text{diag}\{\mathbf{V}^{i_{ss}} : i = 1, \dots, D\}$$

$$= \text{diag}\{\mathbf{g}_{is}^T \Sigma^u \mathbf{g}_{is} + \sigma_e^2 \mathbf{I}_{is} : i = 1, \dots, D\} \quad (4)$$

et

Estimation sur petits domaines sous linéarisation

Hukum Chandra et Ray Chambers¹

Résumé

L'estimation sur petits domaines fondée sur des modèles linéaires mixtes est parfois inefficace quand les relations sous-jacentes ne sont pas linéaires. Nous présentons des techniques d'estimation sur petits domaines pour des variables qui peuvent être modélisées linéairement après une transformation non linéaire. En particulier, nous étendons l'estimateur direct fondé sur un modèle de Chandra et Chambers (2005, 2009) à des données qui concordent avec un modèle linéaire mixte sur l'échelle logarithmique, en utilisant le calage sur un modèle pour définir des poids pouvant être utilisés dans cet estimateur. Nos résultats montrent que l'estimateur fondé sur la transformation que nous obtenons est à la fois efficace et robuste à la distribution des effets aléatoires dans le modèle. Une application à des données d'enquêtes auprès des entreprises démontre la performance satisfaisante de la méthode.

Mots clés : Enquête par sondage ; estimation par sondage ; enquêtes auprès des entreprises ; calage sur un modèle ; données asymétriques ; estimation directe fondée sur un modèle ; meilleur prédicteur sans biais empirique.

1. Introduction

Les méthodes utilisées habituellement pour l'estimation sur petits domaines reposent sur l'hypothèse qu'un modèle linéaire mixte peut être utilisé pour caractériser la relation de régression entre la variable étudiée Y et la variable auxiliaire X dans les petits domaines d'intérêt. En particulier, le meilleur prédicteur linéaire sans biais empirique (EBLUP) (voir Rao, 2003, chapitres 6 à 8) est habituellement fondé sur une hypothèse de modèle linéaire mixte. Toutefois, quand les données sont asymétriques comme cela est souvent le cas dans les enquêtes auprès des entreprises, la relation entre Y et X est parfois non linéaire sur l'échelle originale (brute), mais peut être linéaire sur une échelle transformée, par exemple l'échelle logarithmique (log). Le cas échéant, nous pouvons nous attendre à ce que l'estimation fondée sur un modèle linéaire mixte pour Y soit inefficace comparativement à une estimation fondée sur un modèle similaire pour la version transformée de Y . Voir Hidiroglou et Smith (2005). Le recours à des transformations en inférence n'est pas un fait nouveau (voir, par exemple, Carroll et Ruppert (1988, chapitre 4)). Récemment, Chen et Chen (1996), ainsi que Karlberg (2000a) ont étudié l'utilisation d'une approche de « linéarisation » pour l'estimation par la régression de variables étudiées qui se comportent non linéairement. Cependant, autant que nous sachions, cette idée n'a jamais été appliquée à l'estimation sur petits domaines, même si la théorie économique (et l'observation non formelle) laissent entendre que, dans le cas des données d'enquêtes auprès des entreprises, les relations de régression sont habituellement multiplicatives, et donc linéaires sur l'échelle logarithmique.

Dans le présent article, nous étendons les notions d'estimation directe fondée sur un modèle (DFM) décrites dans Chandra et Chambers (2005, 2009) à la situation où le modèle linéaire mixte qui sous-tend l'estimation sur petits domaines est vérifié sur l'échelle logarithmique, en utilisant des poids calculés par calage sur un modèle (Wu et Sitter 2001). Ce faisant, nous notons que notre approche est facilement généralisable à d'autres transformations monotones (c'est-à-dire inversibles). En revanche, l'extension de l'approche EBLUP à la situation où les données suivent un modèle linéaire mixte sous transformation est compliquée. Nous relaçons également l'hypothèse habituelle de normalité pour les effets de domaine afin d'examiner la robustesse à cette hypothèse.

À la section qui suit, nous résumons l'approche DFM de l'estimation sur petits domaines sous un modèle linéaire mixte. À la section 3, nous décrivons une alternative au modèle linéaire mixte pour des données asymétriques qui se réduit au modèle linéaire mixte sous transformation logarithmique, et à la section 4, nous adoptons une perspective fondée sur un modèle pour motiver l'estimation calée sur un modèle des quantités de populations quand la variable sous-jacente est linéaire après une transformation appropriée. À la section 5, nous regroupons ces deux idées, et introduisons le concept d'un modèle à valeurs prédites dérivé d'un modèle linéaire mixte sous l'échelle transformée. Puis, nous utilisons ce modèle à valeurs prédites pour spécifier les poids de sondage à utiliser dans un estimateur DFM pour l'estimation sur petits domaines. À la section 6, nous présentons les résultats empiriques d'un certain nombre d'études par simulation destinées à comparer l'estimateur DFM sous transformation proposé à l'estimateur

- You et Zhou : Estimation sur petits domaines hiérarchique bayésienne sous un modèle spatial
- You, Y. (2008a). Une approche intégrée de modélisation de l'estimation du taux de chômage pour les régions infraprovinciales au Canada. *Techniques d'enquête*, 34, 21-31.
- You, Y. (2008b). Small area estimation using area level models with model checking and applications. *Proceedings of Survey Methods Section, Statistical Society of Canada*.
- You, Y., et Chapman, B. (2006). Estimation pour petits domaines au moyen de modèles régionaux et d'estimations des variances d'échantillonnage. *Techniques d'enquête*, 32, 107-114.
- You, Y., et Rao, J.N.K. (2000). Estimation bayésienne hiérarchique des moyennes pour petites régions à l'aide de modèles à plusieurs niveaux. *Techniques d'enquête*, 26, 197-206.
- You, Y., et Rao, J.N.K. (2002). Small area estimation using unmatched sampling and linking models. *The Canadian Journal of Statistics*, 20, 3-15.
- Wang, J., et Fuller, W. A. (2003). The mean square error of small area predictors constructed with estimated area variances. *Journal of the American Statistical Association*, 98, 716-723.
- Souza, D.F., Moura, F.A.S. et Migon, H.S. (2009). Prédiction de la population de petits domaines au moyen de modèles hiérarchiques. *Techniques d'enquête*, 35, 221-234.
- Spiegelhalter, D.J., Best, N., Carlin, B.P. and van de Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of Royal Statistical Society*, B, 64, 583-639.
- Sinharay, S., et Stern, H.S. (2003). Posterior predictive model checking in hierarchical models. *Journal of Statistical Planning and Inference*, 111, 209-221.
- Singh, B.B., Shukla, G.K. et Kundu, D. (2005). Modèles spatio-temporels pour l'estimation pour petits domaines. *Techniques d'enquête*, 31, 201-214.

Bibliographie

- Atora, V., et Lahiri, P. (1997). On the superiority of the Bayesian method over the BLUP in small area estimation problems. *Statistica Sinica*, 7, 1053-1063.
- Bayarri, M.J., et Berger, J.O. (2000). P values for composite null models. *Journal of the American Statistical Association*, 95, 1127-1142.
- Béland, Y. (2002). Enquête sur la santé dans les collectivités canadiennes – Aperçu de la méthodologie. Rapports sur la santé, Statistique Canada, n° 82-003 au catalogue, 13, 3, 9-15.
- Besag, J., York, J., et Mollié, A. (1991). Bayesian image restoration, with two applications in spatial statistics (avec discussion). *Annals of the Institute of Statistical Mathematics*, 43, 1-59.
- Best, N., Richardson, S., et Thomson, A. (2005). A comparison of Bayesian spatial models for disease mapping. *Statistical Methods in Medical Research*, 14, 35-39.
- Brown, G., Chambers, R., Heady, P., et Heasman, D. (2001). Evaluation des méthodes d'estimation régionale dans leur application aux estimations du chômage tirées de l'Enquête sur la population active au Royaume-Uni. Recueil : Symposium 2001, *La qualité des données d'un organisme statistique : une perspective méthodologique*, Statistique Canada, CD-ROM, 1-10.
- Chip, S., et Greenberg, E. (1995). Understanding the Metropolitan-Hastings algorithm. *The American Statistician*, 49, 327-335.
- Cressie, N. (1990). Prédiction du sous-dénombrement pour les petites régions à l'aide du modèle linéaire général. Recueil : Symposium 1990, *Mesure et amélioration de la qualité des données*, Statistique Canada, 103-116.
- Daniels, M.J., et Gatsonis, C. (1999). Hierarchical generalized linear models in the analysis of variations in health care utilization. *Journal of the American Statistical Association*, 94, 29-42.
- Datta, G.S., Lahiri, P., Maiti, T., et Lu, K.L. (1999). Hierarchical Bayes estimation of unemployment rates for the states of the U.S. *Journal of the American Statistical Association*, 94, 1074-1082.
- Dick, P. (1995). Modélisation du sous-dénombrement net dans le recensement du Canada de 1991. *Techniques d'enquête*, 21, 51-61.
- Fay, R.E., et Herriot, R.A. (1979). Estimation of income for small places: An application of James-Stein procedures to census data. *Journal of the American Statistical Association*, 74, 268-277.
- Gelfand, A.E. (1996). Model determination using sampling-based methods. Dans *Markov Monte Carlo in Practice* (Eds., W.R. Gilks, S. Richardson et D.J. Spiegelhalter), Londres : Chapman & Hall, 145-161.
- Gelman, A., Carlin, J.B., Stern, H.S. et Rubin, D.B. (2004). *Bayesian Data Analysis*, 2^e Edition. Chapman & Hall/CRC.
- Gelman, A., Meng, X.L. et Stern, H.S. (1996). Posterior predictive assessment of model fitness via realized discrepancies (avec discussion). *Statistica Sinica*, 6, 733-807.
- Gelman, A., et Rubin, D.B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, 7, 457-472.
- Ghosh, M., Natarajan, K., Stroud, T.W.F., et Carlin, B.P. (1998). Generalized linear models for small area estimation. *Journal of the American Statistical Association*, 93, 273-282.
- Ghosh, M., Natarajan, K., Walter, L.A., et Kim, D.H. (1999). Hierarchical Bayes GLMs for the analysis of spatial data: An application to disease mapping. *Journal of Statistical Planning and Inference*, 75, 305-318.
- He, Z., et Sun, D. (2000). Hierarchical Bayes estimation of hunting success rates with spatial correlations. *Biometrics*, 56, 360-367.
- Jiang, J., et Lahiri, P. (2006). Mixed model prediction and small area estimation (avec discussion). *Test*, 15, 1-96.
- Lele, S.R., Dennis, B. et Lutscher, F. (2007). Data cloning: Easy maximum likelihood estimation for complex ecological models using Bayesian Markov chain Monte Carlo methods. *Ecology Letters*, 10, 551-563.
- Lele, S.R., Nadeem, K. et Schumland, B. (2010). Estimability and likelihood inference for generalized linear mixed models using data cloning. *Journal of the American Statistical Association*, 105, 1617-1625.
- Leroux, B.G., Lei, X. et Breslow, N. (1999). Estimation of disease rates in small areas: A new mixed model for spatial dependence. Dans *Statistical Models in Epidemiology, the Environment and Clinical Trials*, (Eds., M.E. Halloran et D. Berry). New York : Springer Verlag, 135-178.
- Liu, B., Lahiri, P. et Kalton, G. (2008). Hierarchical Bayes modeling of survey weighted small area proportions. Manuscript non-publié.
- Maiti, T. (1998). Hierarchical Bayes estimation of mortality rates for disease mapping. *Journal of Statistical Planning and Inference*, 69, 339-348.
- Mohadjer, L., Rao, J.N.K., Liu, B., Krenzke, T. et van de Kerckhove, W. (2007). Hierarchical Bayes small area estimates of adult literacy using unmatched sampling and linking models. *Proceedings of the American Statistical Association, Section of Survey Method Research*.
- Mouna, F.A.S., et Migon, H.S. (2002). Bayesian spatial models for small area proportions. *Statistical Modelling*, 2, 3, 183-201.
- MacNab, Y.C. (2003). Hierarchical Bayesian spatial modeling of small-area rates of non-rare disease. *Statistics in Medicine*, 22, 1761-1773.
- Maiti, T. (1998). Hierarchical Bayes estimation of mortality rates for disease mapping. *Journal of Statistical Planning and Inference*, 69, 339-348.
- Meng, X.L. (1994). Posterior predictive p value. *The Annals of Statistics*, 22, 1142-1160.
- Mollié, A. (1996). Bayesian mapping of diseases. Dans *Markov Chain Monte Carlo in Practice*. Londres : Chapman and Hall, 359-379.
- Rao, J.N.K. (2003). *Small Area Estimation*. New York : John Wiley & Sons, Inc.
- Singh, A.C., Folsom, R.E., Jr. et Vaisish, A.K. (2005). Small area modeling for survey data with smoothed error covariance structure via generalized design effects. Federal Committee on Statistical methods Conference proceedings. Washington, D.C., www.fcsm.gov.
- Statistique Canada, N° 12-001-X au catalogue

Annexe B

Liste des 20 régions sociosanitaires de la Colombie-Britannique
avec les tailles d'échantillon et les structures spatiales correspondantes

Numéro d'ID	Nom de la région sociosanitaire	Taille de l'échantillon	Nombre de régions voisines	Régions voisines
-------------	---------------------------------	-------------------------	----------------------------	------------------

1	East Kootenay	645	3	2, 3, 15
2	West Kootenay-Boundary	705	3	1, 3, 4
3	North Okanagan	890	5	1, 2, 4, 5, 15
4	South Okanagan Similameen	1 063	4	2, 3, 5, 6
5	Thompson	982	7	3, 4, 6, 9, 11, 12, 15
6	Fraser Valley	1 125	5	4, 5, 7, 8, 9
7	South Fraser Valley	1 437	4	6, 8, 17, 19
8	Simon Fraser	1 165	5	6, 7, 9, 17, 18
9	Coast Garibaldi	623	5	5, 6, 8, 11, 18
10	Central Vancouver Island	1 077	2	11, 20
11	Upper Island/Central Coast	746	4	5, 9, 10, 12
12	Cariboo	673	4	5, 11, 13, 15
13	North West	650	3	12, 14, 15
14	Peach Lard	611	2	13, 15
15	Northern Interior	859	6	1, 3, 5, 12, 13, 14
16	Vancouver	1 285	4	17, 18, 19, 20
17	Burnaby	871	5	7, 8, 16, 18, 19
18	North Shore	842	4	8, 9, 16, 17
19	Richmond	828	3	7, 16, 17
20	Capital	1 225	2	10, 16

Nota : Vancouver (n°16) et Capital (n°20) ne sont pas des régions adjacentes sur la carte, puisqu'elles sont séparées par l'océan. Cependant, étant donné le lien intensif et étroit entre ces deux régions, nous les définissons comme voisines dans notre étude à titre d'illustration seulement.

Annexe C

Estimations ponctuelles directes et fondées sur un modèle et CV

ID du domaine	Est. directe	MFH	CAR-MFH	MYC	CAR-MYC
Comparaison des estimations ponctuelles					
1	0,765	0,793	0,812	0,795	0,812
2	0,804	0,795	0,793	0,797	0,794
3	0,745	0,726	0,731	0,725	0,729
4	0,893	0,868	0,874	0,867	0,873
5	0,782	0,739	0,736	0,729	0,731
6	0,943	0,914	0,927	0,918	0,928
7	0,702	0,707	0,712	0,711	0,717
8	0,858	0,845	0,848	0,844	0,849
9	0,877	0,763	0,745	0,765	0,747
10	0,763	0,805	0,799	0,805	0,796
11	0,661	0,685	0,678	0,679	0,676
12	0,717	0,681	0,681	0,680	0,677
13	0,631	0,687	0,692	0,690	0,693
14	0,673	0,685	0,680	0,685	0,686
15	0,793	0,721	0,707	0,728	0,713
16	0,657	0,696	0,702	0,697	0,704
17	0,859	0,778	0,759	0,773	0,759
18	0,583	0,626	0,633	0,618	0,626
19	0,619	0,649	0,647	0,653	0,647
20	0,877	0,923	0,914	0,917	0,908
Comparaison des CV					
1	0,168	0,107	0,099	0,107	0,100
2	0,127	0,105	0,104	0,097	0,093
3	0,135	0,116	0,106	0,110	0,097
4	0,102	0,084	0,076	0,079	0,072
5	0,158	0,094	0,076	0,105	0,083
6	0,113	0,086	0,080	0,086	0,081
7	0,124	0,099	0,096	0,106	0,101
8	0,102	0,085	0,076	0,081	0,073
9	0,158	0,119	0,105	0,117	0,105
10	0,121	0,087	0,086	0,086	0,084
11	0,141	0,118	0,108	0,109	0,105
12	0,196	0,119	0,109	0,130	0,116
13	0,168	0,115	0,108	0,111	0,108
14	0,206	0,126	0,125	0,136	0,133
15	0,121	0,101	0,087	0,094	0,083
16	0,127	0,101	0,097	0,103	0,097
17	0,124	0,107	0,100	0,103	0,096
18	0,155	0,143	0,136	0,134	0,130
19	0,154	0,135	0,134	0,128	0,128
20	0,103	0,086	0,085	0,083	0,082

de recherche de Yong You ont été financées par les ressources de financement global de la recherche de la Direction de la méthodologie de Statistique Canada. Les travaux de Qian M. Zhou ont été exécutés sous la supervision de Yong You durant un stage de recherche à Statistique Canada financé par le MITACS/PNSDC. Q.M. Zhou a présenté les

Annexe A

Lois conditionnelles complètes

A.1. Lois conditionnelles complètes pour l'échantillonnage de Gibbs sous le modèle 1 : MFH.

$$\begin{aligned} & \cdot [\theta_i | y_i, \beta, \sigma_v^2] \sim N[\gamma_i y_i + (1 - \gamma_i) \mathbf{x}_i' \beta, \sigma_v^2 \gamma_i], \text{ où } \gamma_i = \sigma_v^2 / (\sigma_v^2 + \sigma_i^2), \text{ pour } i = 1, \dots, m; \\ & \cdot [\beta | \theta, \sigma_v^2] \sim N \left[\left(\sum_{i=1}^m \mathbf{x}_i \mathbf{x}_i' \right)^{-1} \left(\sum_{i=1}^m \mathbf{x}_i \theta_i \right), \sigma_v^2 \left(\sum_{i=1}^m \mathbf{x}_i \mathbf{x}_i' \right)^{-1} \right]; \\ & \cdot [\sigma_v^2 | \theta, \beta] \sim \text{GI} \left[a_0 + \frac{1}{2} m, b_0 + \frac{1}{2} \sum_{i=1}^m (\theta_i - \mathbf{x}_i' \beta)^2 \right]. \end{aligned}$$

A.2. Lois conditionnelles complètes pour l'échantillonnage de Gibbs sous le modèle 2 : CAR-MFH.

$$\begin{aligned} & \cdot [\theta | y, \beta, \lambda, \sigma_v^2] \sim \text{MVN}(\mathbf{A}y + (\mathbf{I} - \mathbf{A})\mathbf{X}\beta, \mathbf{A}\mathbf{E}), \text{ où } \mathbf{A} = (\mathbf{E}^{-1} + \mathbf{D}/\sigma_v^2)^{-1} \mathbf{E}^{-1} \text{ avec } \mathbf{E} = \text{diag}\{\sigma_1^2, \dots, \sigma_m^2\} \text{ et} \\ & \mathbf{D} = \lambda \mathbf{R} + (1 - \lambda) \mathbf{I}; \end{aligned}$$

$$\begin{aligned} & \cdot [\beta | \theta, \lambda, \sigma_v^2] \sim \text{MVN}[(\mathbf{X}'\mathbf{D}\mathbf{X})^{-1} \mathbf{X}'\mathbf{D}\theta, \sigma_v^2 (\mathbf{X}'\mathbf{D}\mathbf{X})^{-1}]; \\ & \cdot [\lambda | \theta, \beta, \sigma_v^2] \propto |\lambda \mathbf{R} + (1 - \lambda) \mathbf{I}|^{-1/2} \times \exp \left\{ -\frac{1}{2\sigma_v^2} (\theta - \lambda \beta)' [\lambda \mathbf{R} + (1 - \lambda) \mathbf{I}] (\theta - \lambda \beta) \right\}; \\ & \cdot [\sigma_v^2 | \theta, \beta, \lambda] \sim \text{GI} \left[a_0 + \frac{2}{m}, b_0 + \frac{2}{1} (\theta - \lambda \beta)' \mathbf{D} (\theta - \lambda \beta) \right]. \end{aligned}$$

A.3. Lois conditionnelles complètes pour l'échantillonnage de Gibbs sous le modèle 3 : MYC.

$$\begin{aligned} & \cdot [\theta_i | y_i, \beta, \sigma_v^2, \sigma_i^2] \sim N[\gamma_i y_i + (1 - \gamma_i) \mathbf{x}_i' \beta, \sigma_i^2 \gamma_i], \text{ où } \gamma_i = \sigma_v^2 / (\sigma_v^2 + \sigma_i^2), \text{ pour } i = 1, \dots, m; \\ & \cdot [\beta | \theta, \sigma_v^2] \propto N \left[\left(\sum_{i=1}^m \mathbf{x}_i \mathbf{x}_i' \right)^{-1} \left(\sum_{i=1}^m \mathbf{x}_i \theta_i \right), \sigma_v^2 \left(\sum_{i=1}^m \mathbf{x}_i \mathbf{x}_i' \right)^{-1} \right]; \\ & \cdot [\sigma_v^2 | y_i, \theta_i] \sim \text{GI} \left(a_i' + \frac{d_i'}{1} + \frac{2}{(y_i - \theta_i)^2 + d_i'^2}, b_i' + \frac{2}{1} \right), \text{ où } d_i' = n_i - 1, \text{ pour } i = 1, \dots, m; \end{aligned}$$

$$\cdot [\sigma_v^2 | \theta, \beta] \sim \text{GI} \left[a_0 + \frac{1}{2} m, b_0 + \frac{2}{1} \sum_{i=1}^m (\theta_i - \mathbf{x}_i' \beta)^2 \right].$$

A.4. Lois conditionnelles complètes pour l'échantillonnage de Gibbs sous le modèle 4 : CAR-MYC.

$$\begin{aligned} & \cdot [\theta | y, \beta, \lambda, \sigma_v^2, \sigma_i^2] \sim \text{MVN}(\mathbf{A}y + (\mathbf{I} - \mathbf{A})\mathbf{X}\beta, \mathbf{A}\mathbf{E}), \text{ où } \mathbf{A} = (\mathbf{E}^{-1} + \mathbf{D}/\sigma_v^2)^{-1} \mathbf{E}^{-1}, \text{ et} \\ & \mathbf{E} = \text{diag}\{\sigma_1^2, \dots, \sigma_m^2\}, \mathbf{D} = \lambda \mathbf{R} + (1 - \lambda) \mathbf{I}; \\ & \cdot [\beta | \theta, \lambda, \sigma_v^2] \sim \text{MVN}[(\mathbf{X}'\mathbf{D}\mathbf{X})^{-1} \mathbf{X}'\mathbf{D}\theta, \sigma_v^2 (\mathbf{X}'\mathbf{D}\mathbf{X})^{-1}]; \\ & \cdot [\lambda | \theta, \beta, \sigma_v^2] \propto |\lambda \mathbf{R} + (1 - \lambda) \mathbf{I}|^{-1/2} \times \exp \left\{ -\frac{1}{2\sigma_v^2} (\theta - \lambda \beta)' [\lambda \mathbf{R} + (1 - \lambda) \mathbf{I}] (\theta - \lambda \beta) \right\}; \\ & \cdot [\sigma_v^2 | y_i, \theta_i] \sim \text{GI} \left(a_i' + \frac{d_i'}{1} + \frac{2}{(y_i - \theta_i)^2 + d_i'^2}, b_i' + \frac{2}{1} \right), \text{ où } d_i' = n_i - 1, \text{ pour } i = 1, \dots, m; \\ & \cdot [\sigma_v^2 | \theta, \beta, \lambda] \sim \text{GI} \left[a_0 + \frac{2}{m}, b_0 + \frac{1}{2} (\theta - \lambda \beta)' \mathbf{D} (\theta - \lambda \beta) \right]. \end{aligned}$$

avec effets de domaine indépendants à un modèle de corrélation spatiale, et l'avons combiné aux modèles pour petits domaines classiques. Les modèles à corrélation spatiale pour petits domaines CAR-MFH et CAR-MYC que nous proposons comprennent les modèles d'échantillonnage pour petits domaines et un modèle de lien à corrélation spatiale qui reflète l'hétérogénéité non structurée entre les domaines, ainsi que les effets de corrélation spatiale des domaines voisins. Le paramètre d'autocorrélation spatiale ne doit pas être spécifié dans le modèle et est estimé d'après les données.

Dans l'analyse des données, nous avons comparé les modèles spatiaux proposés aux modèles à effets non spatiaux pour estimer le taux de prévalence de l'asthme dans les 20 régions socio-sanitaires de la Colombie-Britannique. Nous constatons que, comparativement aux estimations directes, les estimations fondées sur un modèle représentent une amélioration importante, qui se traduit par des estimations ponctuelles modérément lissées et des CV beaucoup plus petits. En particulier, les modèles proposés sont supérieurs au modèle de Fay-Herriot ou à celui de You-Chapman, que les variances d'échantillonnage soient supposées connues ou inconnues. En outre, la réduction des CV produits par les modèles spatiaux proposés comparativement au modèle de Fay-Herriot ou au modèle de You-Chapman est plus importante pour les domaines ayant un grand nombre de voisins. La comparaison des modèles bayésiens et l'analyse d'adéquation du modèle donnent aussi des résultats favorables aux modèles spatiaux pour petits domaines proposés.

Dans de futurs travaux, les modèles spatiaux pour petits domaines proposés, les modèles spatiaux pour petits domaines proposés pourront être étendus aux modèles d'échantillonnage et de lien non apparés (You et Rao 2002) sous variance d'échantillonnage connue ou inconnue. Nous prévoyons évaluer les effets de divers modèles spatiaux, ainsi que les effets des structures spatiales sur l'estimation. Pour l'analyse des données, nous produirons des estimations de l'état de santé fondées sur les modèles proposés pour les régions socio-sanitaires des diverses régions du Canada et étudierons la possibilité d'étendre l'approche fondée sur un modèle à la production d'estimations à un niveau de détail plus fin, tel que les domaines âge-sexe à l'intérieur des régions socio-sanitaires. Nous prévoyons également examiner la méthode de clonage des données (Lele, Dennis et Lutscher 2007; Lele, Nadeem et Schmuland 2010) pour les modèles spatiaux. Un avantage de cette méthode est que les résultats sont indépendants du choix des priors. Cependant, la demande de ressources informatiques pourrait être considérable.

Remerciements

Nous remercions le rédacteur associé et un examinateur de leurs commentaires et suggestions détaillées. Les travaux

3.5 Diagnostic du biais

Afin d'évaluer le biais éventuel des estimations fondées sur les modèles proposés par rapport aux estimations directes sous le plan de sondage, comme Brown, Chambers, Hoady et Heasman (2001), nous appliquons une simple méthode d'analyse de régression aux estimations directes et aux estimations fondées sur un modèle HB. You (2008a) a également utilisé la méthode d'analyse de régression pour diagnostiquer le biais d'un modèle. Si les estimations fondées sur le modèle sont proches des valeurs réelles des taux de prévalence de la maladie dans les petits domaines, les estimations directes sous le plan de sondage, qui sont supposées fournir des estimations sans biais par rapport aux taux réel de prévalence de la maladie, devraient se comporter comme des variables aléatoires dont les valeurs prédites correspondent aux valeurs des estimations fondées sur un modèle. Autrement dit, les estimations fondées sur un modèle devraient être des prédicteurs sans biais des estimations directes. En ce qui concerne l'analyse de régression, nous ajustons essentiellement le modèle de régression $Y = \alpha + \beta X$ aux données et estimons les coefficients, et nous voyons dans quelle mesure la droite de régression s'approche de $Y = X$. Soit Y les estimations directes et X les estimations fondées sur le modèle. Sous le modèle proposé CAR-MFH, nous obtenons la droite de régression $Y = -0,0021(0,011) + 1,0365(0,1445)X$; sous le modèle proposé CAR-MYC, nous obtenons la droite de régression $Y = -0,0028(0,0108) + 1,0458(0,1427)X$. Donc, les deux droites de régression diffèrent fort peu de $Y = X$. Nous concluons par conséquent que les estimations fondées sur le modèle convergent vers les estimations directes sans biais supplémentaire éventuel induit par les modèles proposés. Les résultats donnent aussi une indication de l'absence de tout biais dû à une erreur éventuelle de spécification du modèle.

4. Conclusion

Dans le présent article, nous avons discuté de deux modèles au niveau du domaine, à savoir le modèle bien connu de Fay-Herriot dans lequel la variance d'échantillonnage est supposée être connue, et le modèle de You-Chapman dans lequel la variance d'échantillonnage est inconnue et modélisée séparément par son estimateur direct. Dans l'un et l'autre modèle, il est supposé que les effets aléatoires de domaine sont des variables aléatoires iid normales pour traduire les effets de l'hétérogénéité inexpliquée des domaines. Après avoir comparé diverses formes de modèles CAR gaussiens proposés dans la littérature (par exemple Best et coll. 2005) pour la cartographie des maladies en vue d'intégrer les effets spatialement corrélés, nous avons étendu le modèle

simple à condition qu'il existe une forme explicite pour la déviance, et p_D peut être calculé après l'exécution de l'échantillonnage de Gibbs en prenant la moyenne d'échantillon des valeurs simulées de $D(\theta)$ dont on soustrait l'estimation de la déviance $D(\hat{\theta})$ obtenue par insertion de valeurs (*plug-in*). Pour les quatre modèles présentés à la section 2, nous avons calculé les valeurs du DIC correspondantes, qui sont présentées au tableau 2. Il est clair que les modèles spatiaux proposés CAR-MFH et CAR-MYC ont tous deux un DIC plus petit que les modèles non spatiaux MFH et MYC, respectivement, ce qui signifie que les modèles spatiaux sont meilleurs que les modèles non spatiaux dans notre étude. Les deux modèles spatiaux CAR-MFH et CAR-MYC donnent de bons résultats dans cet exemple. Les résultats de la comparaison des modèles corroborent les résultats d'estimation présentés à la section 3.2.

Tableau 2
Comparaison des valeurs du DIC pour les quatre modèles hiérarchiques

Modèle	Valeur du DIC
MFH	27,1
CAR-MFH	24,6
MYC	26,8
CAR-MYC	24,5

3.4 Test d'adéquation du modèle

Après de vérifier l'adéquation globale des modèles proposés CAR-MFH et CAR-MYC, nous avons utilisé la méthode de la loi prédictive *a posteriori*. Soit y^{rep} l'observation répétée sous le modèle. La loi prédictive *a posteriori* de y^{rep} sachant les données observées y^{obs} est définie comme étant $f(y^{rep} | y^{obs}) = \int f(y^{rep} | \theta) f(\theta | y^{obs}) d\theta$. Dans cette approche, nous pouvons définir une statistique de test $T(y, \theta)$ qui dépend des données y et éventuellement du paramètre θ , et comparer la valeur observée $T(y^{obs}, \theta | y^{obs})$ à la loi prédictive *a posteriori* de $T(y^{rep}, \theta | y^{obs})$. Le tout écart significatif indiquant l'échec du modèle. Le manque d'ajustement aux données par rapport à la loi prédictive *a posteriori* peut être mesuré par la valeur p de la statistique de test (Meng 1994 ; Gelman, Meng et Stern 1996). La valeur p prédictive *a posteriori* est définie comme étant $p = P(T(y^{rep}, \theta) \geq T(y^{obs}, \theta) | y^{obs})$. Si le modèle donné est bien ajusté aux données observées, $T(y^{obs}, \theta | y^{obs})$ doit être proche de la partie centrale de l'histogramme des valeurs de $T(y^{rep}, \theta | y^{obs})$ si y^{rep} est générée de manière répétée à partir de la loi prédictive *a posteriori*. Par conséquent, la valeur p prédictive *a posteriori* devrait, en principe, être proche de 0,5 si le modèle est adéquatement ajusté aux données. Des valeurs p extrêmes (proche de 0 ou de 1) suggèrent un mauvais ajustement. La vérification des

modèles au moyen de la valeur p prédictive *a posteriori* a été critiquée comme étant trop prudente, à cause de la double utilisation des données observées ; voir, par exemple, Bayarri et Berger (2000). Ces auteurs ont proposé des mesures de la valeur p de rechange pour la vérification des modèles, appelées valeur p prédictive *a posteriori* partielle et valeur p prédictive conditionnelle. Cependant, leurs méthodes sont plus difficiles à mettre en œuvre et à interpréter (Rao 2003 ; Sinharay et Stern 2003). Comme l'ont fait remarquer Sinharay et Stern (2003), la valeur p prédictive *a posteriori* est particulièrement utile si nous considérons le modèle courant comme un point final plausible auquel des modifications ne doivent être apportées que si s'avère que l'adéquation de l'ajustement laisse beaucoup à désirer. Pour exécuter la vérification du modèle prédictif *a posteriori*, nous devons spécifier une statistique de test $T(y, \theta)$. You (2008b) a étudié par simulation plusieurs statistiques de test pour la vérification du modèle prédictif *a posteriori* dans le cas des modèles pour petits domaines et a proposé une statistique de test donnée par

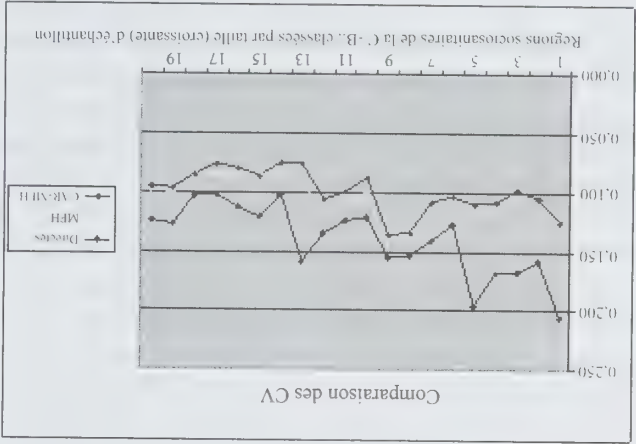
$$T(y, \theta) = |\max(y_i) - \text{moyen}(\theta_i)| - |\min(y_i) - \text{moyen}(\theta_i)|.$$

You (2008b) montre que la statistique de test proposée $T(y, \theta)$ est sensible au choix de la distribution des effets aléatoires et de différentes fonctions de moyenne sous le modèle de Fay-Herriot. Une statistique de test similaire est également proposée dans Gelman et coll. (2004) pour la vérification du modèle prédictif *a posteriori*. Dans notre étude, sous le modèle proposé CAR-MFH, la valeur p estimée est de 0,472, et sous le modèle CAR-MYC, elle est de 0,453. Rien n'indique donc un manque d'ajustement du modèle et les deux modèles spatiaux proposés sont assez bien ajustés aux données.

Pour examiner l'ajustement du modèle au niveau d'observation individuel, nous avons également calculé les valeurs des probabilités prédictives individuelles p_i^* sous la forme $p_i^* = P(y_i^{(rep)} < y_i^{(obs)} | y^{obs})$; voir, par exemple, Gelfand (1996), ainsi que Daniels et Gatsonis (1999). Ces probabilités prédictives individuelles renseignent sur le degré de surestimation ou de sous-estimation systématique des données observées. Pour le modèle CAR-MFH, les p_i^* varient de 0,325 à 0,768 avec une moyenne de 0,517 et une médiane de 0,496 ; pour le modèle CAR-MYC, les p_i^* varient de 0,316 à 0,772 avec une moyenne de 0,511 et une médiane de 0,497. Les deux modèles donnent des résultats fort semblables, et les valeurs moyennes et médianes sont de l'ordre de 0,5. Il n'existe aucune indication d'une surestimation ou d'une sous-estimation systématique des modèles proposés. Les valeurs p globales et les probabilités prédictives individuelles montrent que les modèles spatiaux proposés sont assez bien ajustés aux données.

(sept voisines). La figure montre que les estimations HB d'après le modèle CAR-MFH proposé ont un CV plus petit que celles produites au moyen du modèle de Fay-Herriot. En outre, l'amélioration que représente le modèle CAR-MFH par rapport au modèle de Fay-Herriot est nettement plus marquée dans les régions ayant un grand nombre de voisines, alors que les deux modèles donnent des CV très proches dans les régions dont le nombre de régions adjacentes est plus petit. Nous obtenons des résultats très semblables pour le modèle CAR-MYC comparativement au modèle MYC. Le tableau 1 donne la réduction moyenne des CV pour les régions sociosanitaires ayant le même nombre de voisines. Les résultats du tableau 1 représentent la réduction du CV des modèles spatiaux proposés quand la variance d'échantillonnage est connue ou inconnue. Par exemple, pour σ_i^2 connue (σ_i^2 lissée), pour les régions ne comptant que deux voisines, la réduction moyenne du CV sous le modèle CAR-MFH par rapport au modèle de Fay-Herriot n'est que de 0,9 % environ, tandis que pour les régions comptant sept voisines, la réduction moyenne du CV sous le modèle CAR-MFH par rapport au modèle de Fay-Herriot n'est que de 0,9 % environ, tandis que pour les régions comptant sept voisines, la réduction moyenne du CV sous le modèle CAR-MFH par rapport au modèle MFH peut aller jusqu'à environ 20 %. Pour le cas de σ_i^2 inconnue, nous obtenons des résultats similaires pour CAR-MYC par rapport MYC. Les résultats numériques du tableau 1 confirment la tendance nette à une plus grande réduction des CV sous le modèle spatial proposé que sous le modèle MFH ou MYC à mesure que le nombre de régions voisines augmente. Donc, un plus grand nombre de régions voisines peuvent fournir plus de renseignements sur la structure spatiale en vue d'améliorer la précision et la fiabilité des estimations HB.

Figure 3 CV des estimations directes et HB sous les modèles MFH et CAR-MFH



3.3 Comparaison des modèles bayésiens

À la présente section, nous comparons les modèles proposés CAR-MFH et CAR-MYC aux modèles MFH et MYC, respectivement. Pour la comparaison de modèles hiérarchiques bayésiens, on peut se servir du critère d'information de déviance (DIC pour *Deviance Information Criterion*) proposé par Spiegelhalter, Best, Carlin et van der Linde (2002). Ce critère a été utilisé fréquemment ces dernières années pour comparer les modèles bayésiens à effets non emboîtés et mixtes. Le DIC est basé sur la déviance du modèle $D(\theta)$, qui est égale à moins deux fois la log-vraisemblance du modèle, et il est habituellement calculé sous la forme $DIC = D(\theta) + 2p_D$, où $D(\theta)$ est la déviance du modèle, évaluée à la moyenne *a posteriori* des paramètres du modèle, qui résume la qualité de l'ajustement du modèle, et p_D est le nombre effectif de paramètres, qui traduit la complexité du modèle. p_D est défini comme étant $p_D = \bar{D}(\theta) - D(\hat{\theta})$, où $\bar{D}(\theta)$ est la moyenne *a posteriori* de la déviance du modèle. Donc, le DIC est défini comme la somme de l'adéquation du modèle et de la complexité du modèle. L'ajustement du modèle est d'autant meilleur que la valeur du DIC est faible. Le calcul du DIC est relativement

Figure 4 CV des estimations directes et HB sous les modèles MFH et CAR-MFH avec les régions sociosanitaires classées selon le nombre de régions voisines

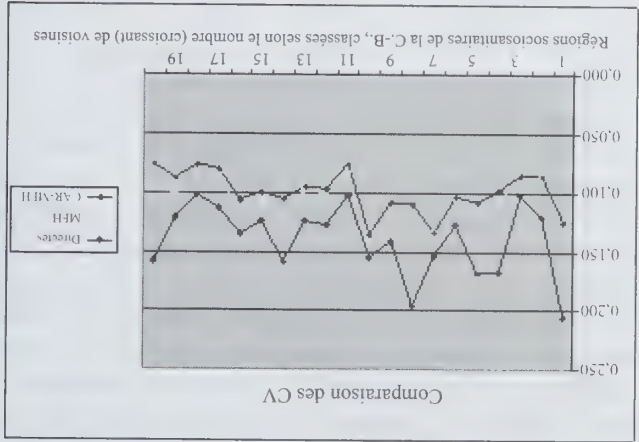


Tableau 1 Comparaison de la réduction moyenne du CV

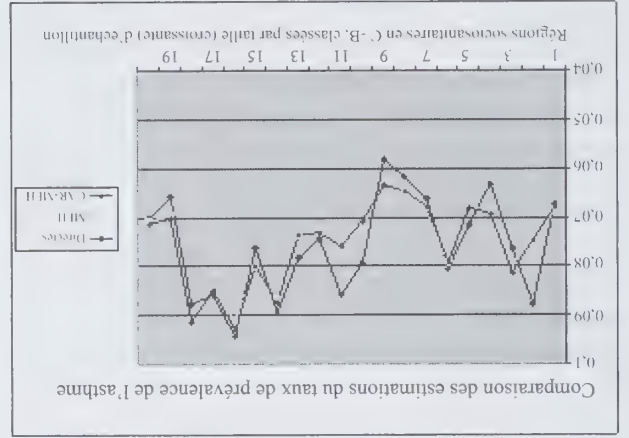
Nombre de régions voisines	Réduction moyenne du CV	
	CAR-MFH par rapport à MFH	CAR-MYC par rapport à MYC
2	0,9 %	1,8 %
3	3,7 %	3,5 %
4	6,3 %	6,0 %
5	8,9 %	8,7 %
6	13,7 %	11,0 %
7	19,2 %	20,7 %

utilisé le prior uniforme $\pi(\lambda) \sim$ pour le paramètre d'autocorrélation. Les priors uniformes sont aussi utilisés fréquemment pour les paramètres d'autocorrélation dans les modèles spatiaux (par exemple, Maiti 1998 ; He et Sun 2000 ; Rao 2003, page 266). Nous avons également essayé plusieurs valeurs différentes pour les priors gamma inverses. Les estimations HB sont assez stables et non sensibles au choix des priors vagues appropriés. Une discussion plus détaillée de l'analyse de sensibilité peut être consultée, par exemple, dans You et Chapman (2006) pour des modèles similaires.

3.2 Comparaison des résultats

Pour commencer, nous présentons les estimations HB du taux de prévalence de l'asthme sous les modèles MFH et CAR-MFH pour les 20 régions socio-sanitaires de la Colombie-Britannique. Les régions socio-sanitaires sont représentées sur l'axe des x, classées par ordre de taille d'échantillon, de la plus petite (Peace Liard) à gauche à la plus grande (South Fraser Valley) à droite. Le modèle 1 (MFH) et le modèle 2 (CAR-MFH) donnent des estimations ponctuelles similaires et les deux estimations fondées sur un modèle mènent à des estimations modérément lisses comparativement aux estimations directes. En outre, les estimations directes et les deux estimations HB du taux de prévalence de la maladie sont très proches pour certaines régions socio-sanitaires dont la taille d'échantillon est grande, mais diffèrent dans une certaine mesure pour certaines régions dont la taille d'échantillon est plus petite. Des résultats similaires sont obtenus sous le modèle 3 (MCY) et le modèle 4 (CAR-MCY).

Figure 2 Estimations directes et fondées sur un modèle HB sous les modèles MFH et CAR-MFH



La figure 3 donne les CV des estimations directes et des deux estimations fondées sur un modèle HB pour les régions socio-sanitaires classées selon la taille d'échantillon, de la plus petite à la plus grande, comme à la figure 2. Les CV des estimations HB sont obtenus en divisant la racine carrée de la variance *a posteriori* par la moyenne *a posteriori*. Comme prévu, les CV des estimations directes ont nettement tendance à diminuer à mesure que la taille d'échantillon augmente. Par contre, les deux estimations fondées sur un modèle HB donnent des CV plus lisses. En outre, ces deux estimations offrent une amélioration importante par rapport aux estimations directes fondées sur le plan en ce qui a trait à la précision et à la fiabilité, c'est-à-dire la réduction des CV. Comparativement aux estimations directes, la réduction moyenne du CV des estimations HB sous le modèle MFH est de l'ordre de 22,7 %, variant de 7,8 % à 40,5 %, et la réduction moyenne du CV des estimations HB sous le modèle CAR-MFH proposé est de 27,8 %, variant de 12,5 % à 52,1 %. Donc, il est clair que le modèle spatial proposé CAR-MFH est supérieur au modèle de Fay-Herriot. Nous avons également obtenu des résultats similaires pour les modèles MYC et CAR-MYC quand la variance d'échantillonnage est modélisée directement. La réduction moyenne du CV sous MYC est de 23,9 %, tandis qu'elle est de 29,0 % sous le modèle spatial proposé CAR-MYC. Des résultats détaillés, y compris les estimations ponctuelles et les CV correspondants, sont présentés dans un tableau à l'annexe C. Dans notre exemple, la taille de l'échantillon au niveau de la région socio-sanitaire est relativement grande. Néanmoins, les estimations fondées sur un modèle révèlent une amélioration importante par rapport aux estimations directes fondées sur le plan de sondage. Nos résultats indiquent que les modèles pour petits domaines proposés peuvent être utilisés afin d'améliorer les estimations directes, même quand la taille d'échantillon est relativement grande. Il convient de souligner que les intervalles de crédibilité bayésiens pour les paramètres de petit domaine peuvent être construits facilement en utilisant la sortie MCMC de l'échantillonneur de Gibbs, si cela est nécessaire en pratique. Il s'agit de l'un des avantages de l'utilisation de l'inférence HB par la voie de l'échantillonnage MCMC. Cependant, ici, nous ne présentons que les estimations ponctuelles fondées sur un modèle et les CV correspondants, car notre objectif principal est de comparer les estimations fondées sur un modèle aux estimations directes et de montrer les gains d'efficacité des modèles. Le gain d'efficacité est clairement appréciable.

Afin d'étudier les effets de l'intégration de la structure spatiale dans le modèle, à la figure 4, nous présentons les CV des estimations directes et HB selon les régions socio-sanitaires classées en fonction du nombre de régions voisines, en allant du plus petit (deux voisines) au plus grand

toutes proches de 1 (inférieures à 1,05), ce qui donne à penser que la convergence souhaitée pour ces paramètres est réalisée par l'échantillonneur de Gibbs.



Figure 1 Carte des 20 régions socioéconomiques de la Colombie-Britannique

Nous avons utilisé des priors vagues pour les hyperparamètres du modèle comme il est d'usage en pratique pour l'estimation sur petits domaines de type HB. En particulier, il est fréquent d'utiliser le prior plat pour le paramètre de régression $\pi(\beta) \propto 1$ et les priors gamma inverses appropriés pour les composantes de la variance (par exemple, Arora et Lahiri 1997; Ghosh et coll. 1998; Datta et coll. 1999; You et Rao 2000; Rao 2003, page 237; Souza et coll. 2009). À l'exemple de MacNab (2003), nous avons

Pour appliquer l'échantillonnage de Gibbs, nous utilisons $L = 5$ exécutions parallèles ayant chacune une longueur de « rodage » de $B = 2\,000$ et une taille d'échantillonnage de Gibbs de $G = 5\,000$. Pour les modèles CAR-MFH et CAR-MYC proposés, afin de réduire l'autocorrélation qui résulte de l'algorithme d'acceptation-rejet dans l'exécution, nous prenons une itération sur cinq après la période de « rodage ». Par conséquent, pour les modèles MFH et MYC, nous avons $n = 5\,000$ échantillons pour chaque exécution, et pour les modèles CAR-MFH et CAR-MYC, nous avons $n = 1\,000$ échantillons pour chaque exécution. Nous surveillons la convergence de l'échantillonnage de Gibbs pour le paramètre de petits domaines θ_i et d'autres paramètres inconnus dans le modèle en utilisant le facteur de réduction potentiel d'échelle (*potential scale reduction factor*) (Gelman et Rubin 1992; Gelman, Carlin, Stern et Rubin 2004, pages 296-297). Nous avons calculé les facteurs de réduction pour tous les paramètres surveillés dans le modèle dans l'échantillonnage de Gibbs. Les valeurs du facteur sont

Soit θ_i le taux réel de prévalence de l'asthme dans la i^{e} région socioéconomique en C.-B., $i = 1, \dots, 20$. D'après les données du cycle 1,1 de l'enquête, nous obtenons l'estimation directe y_i^d de θ_i comme étant le ratio du nombre de personnes asthmatiques (estimation directe) divisé par la taille de population correspondante (constante connue). Nous avons également inclus six variables auxiliaires au niveau du domaine utilisées dans le modèle, à savoir la taille totale de population, le nombre de personnes dont l'asthme est l'un des symptômes de maladie chronique, le nombre de personnes dont l'asthme est le symptôme principal de maladie chronique, le nombre de personnes dont le diabète est l'un des symptômes de maladie chronique, le nombre de personnes dont le diabète est le symptôme principal de maladie chronique, et le nombre de visites à l'hôpital. Notons que, dans la littérature traitant de la cartographie des maladies (par exemple, Mollie 1996; Maiti 1998; MacNab 2003), la loi de Poisson ou la loi binomiale est habituellement celle qui est supposée dans le modèle d'échantillonnage pour l'estimation directe y_i^d . Cependant, dans l'estimation sur petits domaines, l'estimation directe y_i^d s'obtient en se fondant sur le plan d'échantillonnage complexe utilisé dans l'enquête. Donc, on émet habituellement l'hypothèse d'un modèle d'échantillonnage normal pour les estimations directes y_i^d ; voir, par exemple, Datta, Lahiri, Maiti et Lu (1999), Rao (2003), Mohadjer, Rao, Liu, Krenzke et Van de Kerckhove (2007), et You (2008a). Il convient de souligner que nous n'avons pris en considération qu'un seul type de données sur la prévalence de la maladie provenant d'une seule province dans notre étude, et utilisé cet exemple pour illustrer le modèle proposé et évaluer les effets de la modélisation spatiale dans les modèles pour petits domaines.

3. Analyse des données

3.1 Description des données et mise en œuvre

L'Enquête sur la santé dans les collectivités canadiennes (ESCC) est une enquête fédérale réalisée par Statistique Canada. L'objectif principal de l'ESCC est de fournir des estimations à jour et fiables des déterminants de la santé, de l'état de santé et de l'utilisation du système de santé au Canada. Il s'agit d'une enquête transversale dont le cycle de collecte est de deux ans. La première année de l'enquête, le cycle « x.1 » a pour cible les membres de 12 ans et plus de la population à domicile et est une enquête générale sur la santé de la population réalisée auprès d'un grand échantillon (130 000 personnes) conçu en vue de fournir des estimations fiables au niveau des régions sociosanitaires, des provinces et du Canada. La deuxième année de l'enquête, le cycle « x.2 » est réalisé auprès d'un échantillon plus petit (30 000 personnes), réparti en fonction des achats d'unités d'échantillonnage supplémentaires des provinces et conçu pour fournir des résultats au niveau provincial et national sur des thèmes particuliers liés à la santé. Bien que les estimations nationales et provinciales soient très importantes, la demande de données sur la santé à des niveaux plus fins d'aggrégation géographique augmente dans un certain nombre de provinces, dont la Colombie-Britannique (C.-B.), l'Île-du-Prince-Édouard (Î.-P.-É.), le Québec et d'autres. Le cycle « x.1 » de l'ESCC permet de recueillir des données pour 136 régions sociosanitaires réparties dans les dix provinces et les trois territoires. Il est réalisé essentiellement à l'aide de deux bases de sondage. La première, qui est la base principale, est fondée sur la base de sondage aréolaire conçue pour l'Enquête sur la population active du Canada. Cette base aréolaire est utilisée pour échantillonner les logements conformément à un plan d'échantillonnage en grappes stratifiées à plusieurs degrés. La deuxième base de sondage consiste en une liste de numéros de téléphone. La méthode de composition aléatoire est utilisée dans certaines régions sociosanitaires pour des raisons de coût. Des renseignements plus détaillés sur le plan de sondage sont fournis dans Beland (2002). Dans le présent article, nous utilisons un petit ensemble de données provenant du cycle 1.1 pour illustrer l'analyse. Nous voulons estimer le taux de prévalence de la maladie dans les régions sociosanitaires à l'intérieur des provinces. En particulier, nous appliquons les quatre modèles décrits à la section 2 pour estimer le taux de prévalence de l'asthme dans 20 régions sociosanitaires dans la province de la Colombie-Britannique en utilisant les données provenant du cycle 1.1. La figure 1 montre la carte Nous utilisons cette carte pour définir la matrice de corrélation de voisinage utilisée dans les modèles spatiaux. L'annexe B donne la liste des régions sociosanitaires et les structures spatiales connexes.

De nouveau, soulignons que le modèle proposé CAR-MYC se réduit au modèle de You-Chapman quand $\lambda = 0$. Le modèle 3 ainsi que le modèle 4 comportent l'hypothèse implicite que la taille d'échantillon de domaine $n_i \geq 2$. Si des priors plats (β) sont utilisés pour σ_i^2 , il faudrait que $n_i \geq 4$ pour être certains que les lois $a_{posteriori}$ soient appropriées (You et Chapman 2006).

Nous appliquons la méthode d'échantillonnage de Gibbs pour estimer la moyenne $a_{posteriori}$ $E(\theta_i | y)$ et la variance $a_{posteriori}$ correspondante $\text{Var}(\theta_i | y)$. Les lois conditionnelles complètes des paramètres requises sous les divers modèles sont données à l'annexe A. Pour les modèles de Fay-Herriot et de You-Chapman, toutes les lois conditionnelles complètes possèdent des formes explicites et le tirage d'échantillons à partir de ces lois est simple. Pour les modèles spatiaux au niveau du domaine proposés CAR-MFH et CAR-MYC, la loi conditionnelle du paramètre de corrélation spatiale λ ne possède pas de forme explicite. Nous utilisons l'algorithme de Metropolis-Hastings avec l'échantillonneur de Gibbs (Chip et Greenberg 1995) pour mettre à jour λ . Sous le modèle CAR-MFH, la loi conditionnelle complète de λ dans l'échantillonneur de Gibbs peut s'écrire sous la forme

$$[\lambda | \theta, \beta, \sigma_i^2] \propto h(\lambda) f(\lambda)$$

où $f(\lambda)$ est une densité de probabilité de la loi uniforme Uniform(0, 1) donnée par

$$f(\lambda) \propto 1, \text{ où } 0 \leq \lambda \leq 1$$

et $h(\lambda)$ est une fonction donnée par

$$h(\lambda) \propto \left[\lambda \mathbf{R} + (1 - \lambda) \mathbf{I} \right]^{-1} \left\{ \exp \left\{ -\frac{1}{2\sigma_v^2} (\boldsymbol{\theta} - \mathbf{X}\boldsymbol{\beta})' [\lambda \mathbf{R} + (1 - \lambda) \mathbf{I}] (\boldsymbol{\theta} - \mathbf{X}\boldsymbol{\beta}) \right\} \right\}$$

Nous utilisons $f(\lambda)$ comme densité de probabilité génératrice « candidate » dans l'étape de mise à jour de l'algorithme de Metropolis-Hastings. Pour mettre λ à jour à partir des valeurs courantes de $(\boldsymbol{\theta}_{(k)}, \beta_{(k)}, \sigma_{v(2(k))}^2)$, la procédure est la suivante :

1. tirer λ^* d'une loi uniforme ;
2. calculer la probabilité d'acceptation $\alpha(\lambda^*, \lambda_{(k)}) = \min \{ h(\lambda^*) / h(\lambda_{(k)}) , 1 \}$;
3. générer u à partir d'une loi uniforme ; si $u < \alpha(\lambda^*, \lambda_{(k)})$, la valeur candidate λ^* est acceptée, c'est-à-dire $\lambda_{(k+1)} = \lambda^*$; sinon, λ^* est rejetée, et fixée à $\lambda_{(k+1)} = \lambda_{(k)}$.

Pour le modèle CAR-MYC, une procédure similaire peut être appliquée au moment du tirage des échantillons à partir de la loi conditionnelle de λ .

où $w_{i+} = \sum_{j \neq i} w_{ij}$. Le modèle CAR (6) - (7) devient le modèle autorégressif intrinsèque (5) si $\lambda = 1$. Par ailleurs, si $\lambda = 0$, il se réduit au modèle de lien indépendant (1) dans lequel les effets aléatoires propres au domaine v_i sont supposés indépendants. Il convient de souligner que la moyenne et les variances conditionnelles de $b_i | b_{-i}$ sont les sommes pondérées des moments lissés globaux provenant du modèle de lien élémentaire (1) et des moments lissés locaux provenant du modèle autorégressif intrinsèque :

$$E(b_i | b_{-i}) = \frac{1 - \lambda}{1 - \lambda + \lambda w_{i+}} \times 0 + \frac{\lambda w_{i+}}{1 - \lambda + \lambda w_{i+}} \left(\sum_{j \neq i} w_{ij} b_j / w_{i+} \right) + \frac{1 - \lambda + \lambda w_{i+}}{1 - \lambda + \lambda w_{i+}} (\sigma_b^2 / w_{i+}).$$

Donc, le modèle (6) - (7) représente un équilibre entre le modèle de lien indépendant (1) et le modèle CAR intrinsèque (5). Le paramètre de corrélation spatiale λ mesure l'importance des effets spatiaux pour le lissage local des domaines voisins. La structure de modélisation (6) traduit à la fois l'hétérogénéité non structurée entre les domaines et les effets de corrélation spatiale du domaine voisin.

2.3 Modèles hiérarchiques bayésiens et inférence

Afin d'estimer θ_i , le paramètre d'intérêt, nous appliquons une approche hiérarchique bayésienne (HB) en utilisant la méthode d'échantillonnage de Gibbs. Comparativement à d'autres approches, telles que l'EBLUP et l'approche empirique bayésienne (EB), l'approche HB est simple et les inférences concernant θ_i sont exactes contrairement aux inférences EB ou EBLUP. En outre, l'approche HB donne le moyen de traiter des modèles pour petits domaines complexes en utilisant la méthode Monte Carlo par chaîne de Markov (MCMC), qui permet de surmonter en grande partie les difficultés que posent le calcul des intégrales multidimensionnelles des quantités *a posteriori*. Soit $y = (y_1, \dots, y_m)'$, $\theta = (\theta_1, \dots, \theta_m)'$, et $X = (x_1, \dots, x_m)'$. Nous commençons par construire deux modèles HB sans et d'échantillonnage sont connues et remplacées par l'estimation lissée σ_i^2 .

Modèle 1 : Modèle de Fay-Herriot, désigné par MFH (Fay et Herriot 1979 ; Rao 2003).

- $y_i | \theta_i \sim N(\theta_i, \sigma_i^2 = \sigma_y^2)$, pour $i = 1, \dots, m$;
- $\theta_i | \beta, \sigma_y^2 \sim N(x_i' \beta, \sigma_y^2)$, pour $i = 1, \dots, m$;

Il convient de souligner que le modèle CAR-MFH proposé se réduit au modèle MFH lorsque le paramètre d'autocorrélation spatial est $\lambda = 0$. Nous considérons aussi deux modèles HB dont la variance d'échantillonnage σ_i^2 est inconnue et modélisée par l'estimateur sans biais direct s_i^2 .

Modèle 3 : Modèle de You-Chapman désigné par MYC (You et Chapman 2006).

- $y_i | \theta_i, \sigma_i^2 \sim N(\theta_i, \sigma_i^2)$, pour $i = 1, \dots, m$;
- $d_i s_i^2 | \sigma_i^2 \sim \sigma_i^2 \chi_{d_i}^2$ où $d_i = n_i - 1$, pour $i = 1, \dots, m$;
- $\theta_i | \beta, \sigma_y^2 \sim N(x_i' \beta, \sigma_y^2)$, pour $i = 1, \dots, m$;
- Priors pour les paramètres inconnus $(\beta, \sigma_y^2, \sigma_i^2, i = 1, \dots, m)$: $\pi(\beta) \propto 1$; $\pi(\sigma_y^2) \sim \text{GI}(a_0, b_0)$; $\pi(\sigma_i^2) \sim \text{GI}(a_i, b_i)$ pour $i = 1, \dots, m$, où a_i, b_i ($0 \leq i \leq m$) sont des constantes connues choisies très petites pour refléter les connaissances vagues au sujet de σ_i^2 et σ_y^2 .

Modèle 4 : Modèle CAR au niveau du domaine proposé avec variances d'échantillonnage inconnues, en tant qu'extension du modèle de You-Chapman, désigné par CAR-MYC.

- $y | \theta, \sigma_y^2, \sigma_i^2 \sim \text{MVN}(\theta, E)$, où la matrice E contient les éléments diagonaux σ_i^2 ;
- $d_i s_i^2 | \sigma_i^2 \sim \sigma_i^2 \chi_{d_i}^2$ où $d_i = n_i - 1$, pour $i = 1, \dots, m$;
- $\theta | \beta, \sigma_y^2 \sim \text{MVN}(X\beta, \sigma_y^2 D^{-1})$, où $D = \lambda R + (1 - \lambda) I$;
- priors pour les paramètres $(\beta, \lambda, \sigma_y^2, \sigma_i^2, i = 1, \dots, m)$: $\pi(\beta) \propto 1$; $\pi(\lambda) \sim \text{Uniform}(0, 1)$, où $0 \leq \lambda \leq 1$; $\pi(\sigma_y^2) \sim \text{GI}(a_0, b_0)$; $\pi(\sigma_i^2) \sim \text{GI}(a_i, b_i)$ pour $i = 1, \dots, m$, où a_i, b_i ($0 \leq i \leq m$) sont des constantes connues choisies très petites.

gamma inverse.

- Priors pour les paramètres (β, σ_y^2) : $\pi(\beta) \propto 1$; $\pi(\sigma_y^2) \sim \text{GI}(a_0, b_0)$, où a_0, b_0 sont des constantes connues choisies très petites pour refléter les connaissances vagues au sujet de σ_y^2 . N désigne la loi normale et GI , la loi gamma inverse.

Modèle 2 : Modèle CAR au niveau du domaine proposé, en tant qu'extension du modèle de Fay-Herriot, désigné par CAR-MFH.

- $y | \theta \sim \text{MVN}(\theta, E)$, où E est une matrice diagonale dont le i^{e} élément diagonal est $\sigma_i^2 = \sigma_y^2$;
- $\theta | \beta, \sigma_y^2 \sim \text{MVN}(X\beta, \sigma_y^2 D^{-1})$, où $D = \lambda R + (1 - \lambda) I$, avec I désignant une matrice identité de dimension m et R , la matrice de voisinage ;
- Priors pour les paramètres $(\beta, \lambda, \sigma_y^2)$: $\pi(\beta) \propto 1$; $\pi(\lambda) \sim \text{Uniform}(0, 1)$, où $0 \leq \lambda \leq 1$; $\pi(\sigma_y^2) \sim \text{GI}(a_0, b_0)$, où a_0, b_0 sont des constantes connues choisies très petites. MVN désigne la loi normale multivariée.

Il convient de souligner que le modèle CAR-MFH proposé se réduit au modèle MFH lorsque le paramètre d'autocorrélation spatial est $\lambda = 0$.

Nous considérons aussi deux modèles HB dont la variance d'échantillonnage σ_i^2 est inconnue et modélisée par l'estimateur sans biais direct s_i^2 .

inconnue. En pratique, un choix fréquent pour w_{ij} consiste à poser que $w_{ij} = 0$ à moins que les domaines i et j soient voisins (c'est-à-dire aient une limite commune), auquel cas $w_{ij} = 1$. Le modèle (4) est proposé par Besag, York et Mollié (1991) pour séparer les effets spatiaux de l'hétérogénéité globale dans les domaines. Dans le modèle (4), les effets aléatoires indépendants v_i traduisent l'hétérogénéité géographique non structurée entre les domaines, et les effets aléatoires spatiaux u_i traduisent la dépendance spatiale entre les domaines. De cette façon, le degré de dépendance spatiale globale peut être exprimé en se basant sur la proportion de la variation totale dans les $v_i + u_i$ reflétées par chaque composante.

En pratique, le choix entre un modèle non structuré (par exemple, le modèle de lien élémentaire) donné par (1) et un modèle purement structuré spatialement (par exemple, le modèle autorégressif intrinsèque) donné par (5) n'est souvent pas clair. Pour le modèle (4), l'inférence *a posteriori* au sujet de la dépendance spatiale est fondée sur la proportion de la variation totale de la somme de $v_i + u_i$ reflétée par chaque composante. Cependant, bien que les lois conditionnelles univariées de la composante spatiale (5) soient bien définies, la loi conjointe correspondante est impropre (de moyenne non définie et de variance infinie). De surcroît, le modèle (4) pose un problème éventuel d'identifiabilité où seule la somme des effets aléatoires $v_i + u_i$ est bien définie par les données ; voir, par exemple, Best et coll. (2005), pour une discussion plus détaillée.

Nous pouvons également considérer une autre paramétrisation spatiale étudiée par Leroux, Lei, et Breslow (1999) et par MacNab (2003), qui permet d'éviter le problème d'identifiabilité qui se pose avec le modèle (4). Soit $\theta_i = x_i'\beta + b_i$, et $\mathbf{b} = (b_1, \dots, b_m)'$. À l'instar de Leroux et coll. (1999) et de MacNab (2003), nous appliquons le modèle conditionnel autorégressif (CAR) qui suit aux effets spatiaux de domaine $\mathbf{b} = (b_1, \dots, b_m)'$:

$$\mathbf{b} \sim \text{MVN}(\mathbf{0}, \Sigma(\sigma_b^2, \lambda)) \quad (6)$$

$$\Sigma(\sigma_b^2, \lambda) = \sigma_b^2 \mathbf{D}^{-1}, \quad \mathbf{D} = \lambda \mathbf{R} + (\mathbf{I} - \lambda) \mathbf{I} \quad (7)$$

où σ_b^2 est un paramètre de dispersion spatiale et λ est un paramètre d'autocorrélation spatiale, $0 \leq \lambda \leq 1$; \mathbf{I} est une matrice identité de dimension m ; \mathbf{R} est une matrice, habituellement appelée matrice de voisinage (*neighbourhood matrix*), dont le i^{e} élément diagonal est égal au nombre de voisins du domaine i , et dont les éléments hors diagonale dans chaque ligne sont égaux à -1 si les domaines correspondants sont voisins et à 0 autrement. Le modèle CAR donné par les expressions (6) et (7) correspond à la loi conditionnelle de b_i suivante :

$$b_i | b_{-i} \sim N \left(\frac{\lambda}{1 - \lambda + \lambda w_{ii}} \sum_{j \neq i} w_{ij} b_j, \frac{\sigma_b^2}{1 - \lambda + \lambda w_{ii}} \right),$$

années, une méthode de lissage des effets de plan a été élaborée et utilisée en pratique pour obtenir des estimateurs de variance lissés (par exemple, Singh, Folsom et Vaish 2005 ; You 2008a ; Liu, Lahiri et Kalton 2008). En particulier, You (2008a) a appliqué une approche de modélisation à effets de plan égaux pour obtenir des estimations lissées des variances d'échantillonnage. L'effet de plan pour le i^{e} domaine peut s'écrire approximativement sous la

forme

$$\text{deff}_i = \frac{s_i^2}{S_i^2}, \text{ pour } i = 1, \dots, m,$$

où s_i^2 est l'estimation directe sans biais de la variance d'échantillonnage fondée sur le plan d'échantillonnage complexe, et S_i^2 est l'estimation de la variance d'échantillonnage sous un plan d'échantillonnage aléatoire simple. Pour chaque domaine, en émettant l'hypothèse d'un effet de plan commun, un facteur deff lissé peut être obtenu en appliquant deff = $\sum_{i=1}^m \text{deff}_i / m$. Ensuite, une estimation lissée de la variance d'échantillonnage $\tilde{\sigma}_i^2$ peut être obtenue sous la forme $\tilde{\sigma}_i^2 = s_i^2 \cdot \text{deff}_i$.

Au lieu d'introduire les estimations lissées des variances d'échantillonnage dans le modèle, nous pouvons modéliser directement ces variances. Dans Wang et Fuller (2003) et You et Chapman (2006), ces auteurs supposent que la variance d'échantillonnage σ_i^2 est inconnue et estiment σ_i^2 au moyen d'un estimateur direct sans biais s_i^2 , qui est indépendant de l'estimateur direct sous le plan y_i . Ils supposent aussi que $d_i s_i^2 \sim \sigma_i^2 \chi_{d_i}^2$, où $d_i = n_i - 1$ et n_i est la taille d'échantillon pour le i^{e} domaine. You et Chapman (2006) ont examiné l'approche HB complète avec la méthode d'échantillonnage de Gibbs qui tient compte automatique de l'incertitude supplémentaire associée à l'estimation de σ_i^2 . Dans le présent article, nous considérons à la fois les approches de lissage et de modélisation pour les variances d'échantillonnage.

2.2 Modèles spatiaux

En vue d'intégrer les effets spatiaux spatialement corrélés dans le modèle de lien, un moyen simple et évident consiste à ajouter un effet aléatoire spatial u_i dans le modèle de lien indépendant (1) comme il suit :

$$\theta_i = x_i'\beta + v_i + u_i, \quad (4)$$

où les u_i suivent le modèle conditionnel autorégressif (CAR) intrinsèque bien connu donné par

$$u_i | u_{-i} \sim N \left(\frac{\sum_{j \neq i} w_{ij} u_j}{\sum_{j \neq i} w_{ij}}, \frac{\sigma_u^2}{\sum_{j \neq i} w_{ij}} \right), \quad (5)$$

où u_i désigne les valeurs des effets aléatoires spatiaux u_i dans tous les autres domaines avec $j \neq i$, les poids w_{ij} sont des constantes fixées et σ_u^2 est une composante de variance

proposés au modèle de Fay-Herriot et au modèle de You-Chapman (You et Chapman 2006) afin d'étudier les effets de la prise en compte de la structure spatiale sur les effets aléatoires de domaine. Nous présentons aussi une comparaison des modèles bayésiens et une analyse de l'adéquation du modèle. Enfin, à la section 5, nous tirons certaines conclusions.

2. Modèles et inférence pour petits domaines

2.1 Modèle de Fay-Herriot

Soit θ_i le paramètre d'intérêt pour le i^{e} domaine, où $i = 1, \dots, m$, et m est le nombre total de domaines. Le modèle de Fay-Herriot repose sur l'hypothèse que les θ_i sont reliés à des données auxiliaires particulières au domaine $\mathbf{x}_i = (x_{i1}, \dots, x_{ip})'$ au moyen du modèle de régression linéaire qui suit :

$$\theta_i = \mathbf{x}_i' \boldsymbol{\beta} + v_i, \quad i = 1, \dots, m \quad (1)$$

où $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)'$ est le vecteur de dimension $p \times 1$ des coefficients de régression, et les v_i sont les effets aléatoires propres au domaine que l'on suppose être *iid* avec $E(v_i) = 0$ et $\text{Var}(v_i) = \sigma_v^2$. L'hypothèse de normalité peut également être incluse. Ce modèle est appelé modèle de lien pour θ_i . Le modèle de Fay-Herriot repose aussi sur l'hypothèse qu'un estimateur direct fondé sur le plan y_i , qui est habituellement sans biais sous le plan pour le paramètre d'intérêt θ_i , existe quand la taille d'échantillon du domaine $n_i > 1$. On suppose habituellement que

$$y_i = \theta_i + e_i, \quad i = 1, \dots, m \quad (2)$$

où les e_i sont les erreurs d'échantillonnage associées à l'estimateur direct y_i . Nous supposons en outre que les e_i sont des variables aléatoires normales indépendantes de moyenne $E(e_i | \theta_i) = 0$ et de variance d'échantillonnage $\text{Var}(e_i | \theta_i) = \sigma_e^2$. Le modèle (2) est appelé modèle d'échantillonnage pour l'estimateur direct y_i . La combinaison des deux composantes (1) et (2) donne un modèle linéaire à effets mixtes, le modèle de Fay-Herriot, de la forme

$$y_i = \mathbf{x}_i' \boldsymbol{\beta} + v_i + e_i, \quad i = 1, \dots, m. \quad (3)$$

Dans le modèle de Fay-Herriot élémentaire (3), on suppose habituellement que les variances d'échantillonnage σ_e^2 sont connues, ce qui est une hypothèse très forte. En général, nous pouvons utiliser les estimations directes des variances d'échantillonnage d'après les données d'enquête, mais ces estimations sont instables si les tailles d'échantillon sont faibles. Par conséquent, en pratique, on introduit dans le modèle un estimateur lissé de σ_e^2 que l'on traite alors comme étant connue. La fonction de variance généralisée est habituellement appliquée en pratique pour obtenir un estimateur lissé de la variance d'échantillonnage (par exemple, Dick 1995). Ces dernières

d'échantillonnage que l'on traite ensuite comme étant connue dans le modèle. Wang et Fuller (2003), ainsi que You et Chapman (2006) ont étudié la situation où les variances d'échantillonnage sont inconnues et modélisées séparément par des estimateurs directs. Dans le présent article, nous examinerons les méthodes de lissage et de modélisation pour les variances d'échantillonnage dans le modèle d'échantillonnage.

Le modèle de lien relie le paramètre d'intérêt à un modèle de régression contenant des effets aléatoires propres au domaine. Dans le modèle de Fay-Herriot, on suppose habituellement que les effets aléatoires de domaine sont des variables aléatoires indépendantes et identiquement distribuées (*iid*) qui suivent une loi normale pour refléter les variations géographiquement non structurées entre les domaines. Cependant, dans certaines applications aux petits domaines, particulièrement dans les problèmes d'estimation de la santé publique, la variation géographique de la prévalence d'une maladie est un sujet d'intérêt, et l'estimation des profils spatiaux globaux de risque et l'emprunt d'information à diverses régions pour réduire les variances des estimations finales sont deux éléments importants. Donc, il pourrait être plus raisonnable de construire des modèles spatiaux sur les effets aléatoires propres aux domaines pour traduire l'interdépendance spatiale de ces effets. Les modèles spatiaux sont généralement utilisés pour l'estimation sur petits domaines ayant trait à la santé, et plusieurs de ces modèles ont été proposés (par exemple, Cressie 1990; Ghosh, Natarajan, Stroud et Carling 1998; Maïti 1998; Ghosh, Natarajan, Walter et Kim 1999; He et Sun 2000; Moura et Migon 2002; Singh, Shukla et Kundu 2005; Souza, Moura et Migon 2009). Best, Richardson et Thomson (2005) ont présenté une revue complète des modèles spatiaux pour la cartographie des maladies. Rao (2003) a également discuté de plusieurs modèles spatiaux pour petits domaines.

L'objectif du présent article est d'examiner des modèles à corrélation spatiale pour l'estimation sur petits domaines et d'illustrer leur utilité en les appliquant à des données d'enquête sur la santé. La présentation de l'article est la suivante. À la section 2, nous commençons par étudier les modèles au niveau du domaine, y compris le modèle de Fay-Herriot et les modèles de lien à corrélation spatiale. Ensuite, à la section 3, nous proposons des modèles d'estimation sur petits domaines hiérarchiques bayésiens (HB) avec corrélation spatiale et obtenons l'inférence HB pour les paramètres de petit domaine grâce à la méthode d'échantillonnage de Gibbs. À la section 4, nous appliquons les modèles proposés à l'analyse de données sur des petits domaines provenant de l'Enquête sur la santé dans les collectivités canadienne. Nous comparons la performance des estimations fondées sur un modèle aux estimations directes fondées sur le plan et, en outre, nous comparons les modèles

Estimation sur petits domaines hiérarchique bayésienne sous un modèle spatial avec application à des données d'enquête sur la santé

Yong You et Qian M. Zhou¹

Résumé

Dans le présent article, nous étudions l'estimation sur petits domaines en nous servant de modèles au niveau du domaine. Nous considérons d'abord le modèle de Fay-Herriot (Fay et Herriot 1979) pour le cas d'une variance d'échantillonnage connue lissée et le modèle de You-Chapman (You et Chapman 2006) pour le cas de la modélisation de la variance d'échantillonnage. Ensuite, nous considérons des modèles spatiaux hiérarchiques bayésiens (HB) qui étendent les modèles de Fay-Herriot et de You-Chapman en tenant compte à la fois de l'hétérogénéité géographique non structurée et des effets de corrélation spatiale entre les domaines pour le lissage local. Les modèles proposés sont mis en œuvre en utilisant la méthode d'échantillonnage de Gibbs pour une inférence entièrement bayésienne. Nous appliquons les modèles proposés à l'analyse de données d'enquête sur la santé et comparons les estimations fondées sur le modèle HB aux estimations directes fondées sur le plan. Nos résultats montrent que les estimations fondées sur le modèle HB ont de meilleures propriétés que les estimations directes. En outre, les modèles spatiaux au niveau du domaine proposés produisent des CV plus petits que les modèles de Fay-Herriot et de You-Chapman, particulièrement pour les domaines ayant trois domaines voisins ou plus. Nous présentons aussi une comparaison des modèles bayésiens et une analyse de l'adéquation du modèle.

Mots clés : Modèle au niveau du domaine ; comparaison de modèles bayésiens ; taux de prévalence de la maladie ; échantillonnage de Gibbs ; modèle spatial hiérarchique ; vérification du modèle prédicatif *a posteriori* ; variance d'échantillonnage.

1. Introduction

Les méthodes d'estimation sur petits domaines fondées sur un modèle ont été utilisées à grande échelle en pratique en raison de la demande croissante d'estimations précises pour des régions locales et divers petits domaines. En général, les enquêtes par sondage sont conçues pour fournir des estimations fiables pour de grandes régions ou des agrégats de petits domaines, tel que le pays dans son ensemble et les provinces. Des estimations par sondage directes fondées uniquement sur des données d'échantillon propres au domaine fournissent habituellement des estimations fiables du paramètre d'intérêt pour les grands domaines. Pour les petits domaines, particulièrement certaines petites régions géographiques ou des petits domaines particuliers, les estimations directes sont susceptibles de produire de grandes erreurs-types, à cause de la petite taille des échantillons dans ces petits domaines. Par conséquent, en inférence pour les petits domaines, il est nécessaire d'emprunter de l'information à des domaines connexes pour produire des estimations indirectes qui augmentent les tailles effectives d'échantillons et donc, la précision des estimations. Il est généralement reconnu aujourd'hui que ces estimations indirectes doivent être fondées sur des modèles explicites qui fournissent des liens avec les domaines connexes grâce à l'utilisation de données supplémentaires, telles que des chiffres de recensement ou des données de dossiers administratifs ; voir, par exemple, Rao (2003) et

Jiang et Lahiri (2006) pour une discussion plus approfondie des méthodes d'estimation sur petits domaines fondée sur un modèle. Les estimations fondées sur un modèle sont obtenues en vue d'améliorer les estimations directes fondées sur le plan en ce qui a trait à la précision et à la fiabilité, c'est-à-dire réduire les coefficients de variation (CV). Ces modèles d'estimation sur petits domaines se répartissent en deux grandes catégories, à savoir les modèles au niveau du domaine et les modèles au niveau de l'unité. Les modèles au niveau du domaine sont fondés sur des estimations directes au niveau du domaine et les modèles au niveau de l'unité, sur des observations individuelles faites dans les petits domaines. Dans le présent article, nous nous concentrons sur les modèles au niveau du domaine qui empruntent de l'information à diverses régions pour améliorer les estimations directes sous le plan.

Parmi les modèles au niveau du domaine, celui de Fay-Herriot (Fay et Herriot 1979) est un modèle élémentaire très utilisé en pratique pour obtenir des estimations fondées sur un modèle fiables pour de petits domaines. Fondamentalement, le modèle de Fay-Herriot possède deux composantes, à savoir un modèle d'échantillonnage pour les estimations directes et un modèle de lien pour les paramètres d'intérêt. Le modèle d'échantillonnage comprend l'estimation par sondage direct et la variance d'échantillonnage correspondante. Le modèle de Fay-Herriot repose sur l'hypothèse que la variance d'échantillonnage est connue dans le modèle. Habituellement, on obtient un estimateur lissé de la variance

- Chipperfield, Bishop et Campbell : Estimation du maximum de vraisemblance pour les tableaux de contingence
- Winkler, W.E. (2005). Approximate String Comparator Search Strategies for Very Large Administrative Lists. Collection de rapports de recherche statistique, n° RRS2005/02, Bureau of the Census.
- Wright, J., Bishop, G. et Ayre, T. (2009). Assessing the Quality of Linking Migrant Settlement Records to Census Data. Travaux de recherche de méthodologie, n° de catalogue 1351.0.55.027, Australian Bureau of Statistics, Canberra.

En général, traiter naïvement un fichier de données appariées comme si l'appariement était parfait donne lieu à des estimations biaisées. L'analyste ne devrait adopter l'approche naïve que si le nombre d'enregistrements non appariés, définis comme étant des enregistrements qui pourraient être appariés correctement, mais qui ne l'ont pas été du tout, ainsi que le nombre d'appariements incorrects sont négligeables. Le présent article décrit une approche fondée sur le maximum de vraisemblance en vue de faire de inférences valides en présence des deux sources d'erreur. Cette approche s'appuie sur l'algorithme EM bien connu et est facile à appliquer en pratique. La méthode peut être utilisée quand l'un des fichiers n'est pas nécessairement un sous-ensemble de l'autre et que l'appariement comporte des passages multiples. Ces situations se présentent fréquemment en pratique, comme en témoignent de nombreux exemples récents à l'Australian Bureau of Statistics. L'étude empirique montre que l'approche du MV améliore de manière significative les estimations fondées sur des données appariées.

Dans le cas particulier où le fichier X est obtenu par tirage d'un échantillon aléatoire du fichier Y, la procédure d'estimation décrite n'est pas « complètement » celle du maximum de vraisemblance, parce qu'elle ne s'appuie pas sur le fait que les totaux de population pour le fichier Y sont connus. Bien que l'inférence selon la méthode décrite ici demeure valide dans ce cas, il y aurait peut-être moyen de la rendre plus efficace (voir Scott et Wild 1997).

Les auteurs remercient Raymond Chambers et deux examinateurs de *Techniques d'enquête* de leur contribution au présent article.

Bibliographie

- Chambers, R., Chipperfield, J.O., Davis, W. et Kovacevic, M. (2009). Regression Inference Based on Estimating Equations and Probability-Linked Data. Soumis pour publication.
- Chambers, R.L., et Skinner, C.J. (2003). *Analysis of Survey Data*. New York : John Wiley & Sons, Inc.
- Bishop, G. (2009). Assessing the Likely Quality of the Statistical Longitudinal Census Dataset. Travaux de recherche de méthodologie, n° de catalogue 1351.0.55.026. Australian Bureau of Statistics, Canberra.
- Australian Bureau of Statistics (2008). Census Data Enhancement - Indigenous Mortality Quality Study. 2006-07. Document d'information n° de catalogue 4723.0.
- Bishop, G. (2009). Assessing the Likely Quality of the Statistical Longitudinal Census Dataset. Travaux de recherche de méthodologie, n° de catalogue 1351.0.55.026. Australian Bureau of Statistics, Canberra.
- Chambers, R., et Bishop, G. (2009). A Linkage Method for the Formation of the Statistical Longitudinal Census Dataset. Travaux de recherche de méthodologie, n° de catalogue 1351.0.55.025, Australian Bureau of Statistics, Canberra.
- Winkler, W.E. (2001). Record Linkage Software and Methods for Merging Administrative Lists. Collection de rapports de recherche statistique, n° RR2001/03, Bureau of the Census.
- Statistique Canada, N° 12-001-X au catalogue
- Chambers, R. (2008). Regression analysis of probability-linked data. *Statisphere*, Volume 4, <http://www.statisphere.govt.nz/official-statistics-research/series/vol-4.htm>.
- Christen, P., et Churches, T. (2005). Febtl – Freely extensible biomedical record linkage. Version 0.3.1. vue le 17 novembre 2008. <http://cs.anu.edu.au/~Peter.Christen/Febtl/febtl-0.3.febtl-doc-0.3/contents.html>.
- Conn, L., et Bishop, G. (2006). Exploring Methods for Creating a Longitudinal Census Dataset. Articles du comité consultative sur la méthodologie, n° de catalogue 1352.0.55.076, Australian Bureau of Statistics, Canberra.
- Fair, M. (2004). Generalized record linkage system-Statistics Canada's record linkage software. *Austrian Journal of Statistics*, 33(1 et 2), 37-53.
- Fellegi, I.P., et Sunter, A.B. (1969). A theory for record linkage. *Journal of the American Statistical Association*, 64, 1183-1210.
- Fuller, W.A. (1987). *Measurement Error Models*. New York : John Wiley & Sons, Inc.
- Hausman, J.A., Abrevaya, J. et Scott-Morton, F.M. (1998). Misclassification of the dependent variable in a discrete-response setting. *Journal of Econometrics*, 87, 239-269.
- Herzog, T.N., Scheuren, F.J. et Winkler, W.E. (2007). *Data Quality and Record Linkage Techniques*. New York : Springer.
- Holman, C.D.J., Bass, A.J., Rouse, I.L. et Hobbs, M.S.T. (1999). Population-based linkage of health records in Western Australia: Development of a health services research linked database. *Australian and New Zealand Journal of Public Health*, 23(5), 453-459.
- Lahiri, P., et Larsen, M.D. (2005). Regression analysis with linked data. *Journal of the American Statistical Association*, 100, 222-230.
- National Center for Health Statistics (2009). Linkages between Survey Data from the National Center for Health Statistics and Program Data from the Social Security Administration. Rapport de méthodologie, http://www.cdc.gov/nchs/data/datalinkage/ssa_methods_report_2009.pdf.
- Rubin, D.B., et Little, R.J.A. (2003). *Statistical analysis of missing data*, 2^e Edition. New York : John Wiley & Sons, Inc.
- Scheuren, F., et Winkler, W.E. (1993). Analyse de régression de fichiers de données couplés par ordinateur. *Techniques d'enquête*, 19, 45-65.
- Scott, A.J., et Wild, C.J. (1997). Fitting regression models to case-control data by maximum likelihood. *Biometrika*, 84, 57-71.
- Solon, R., et Bishop, G. (2009). A Linkage Method for the Formation of the Statistical Longitudinal Census Dataset. Travaux de recherche de méthodologie, n° de catalogue 1351.0.55.025, Australian Bureau of Statistics, Canberra.

Tableau 3

Pourcentages de l'ensemble des personnes de plus de 15 ans dans les diverses catégories d'emploi en 2005. Pour chaque ensemble de données appartenant à la section 3.2, et de la méthode naïve, qui repose sur l'hypothèse qu'il n'existe aucune erreur d'appariement,

Estimations pour divers ensembles de données appartenant à la norme d'or

Situation en 2006	or	Faible	
		Naïve	MV

a : Personnes occupées en 2005	or	Faible	
		Naïve	MV
Occupé	91,8	92,2	92,6
En chômage	1,8	1,7	1,9
Inactif	6,2	6,1	5,6

b : Personnes en chômage en 2005	Occupé	Faible	
		Naïve	MV
Occupé	44,5	44,3	44,0
En chômage	26,8	26,6	27,5
Inactif	28,6	28,7	28,4

c : Personnes inactives en 2005	Occupé	Faible	
		Naïve	MV
Occupé	12,1	12,3	11,1
En chômage	3,1	3,0	3,0
Inactif	84,7	84,5	85,7

Tableau 4

Situation d'étudiant en 2006 pour les étudiants du secondaire en 2005

Situation d'étudiant en 2006	Or	Faible	
		Naïve	MV
Élève du secondaire	79,3	79,3	79,6
Études secondaires terminées	14,0	14,3	13,7
non terminées	6,6	6,3	6,6

5.3 Simulation

L'étude par simulation qui suit illustre les problèmes que pose l'analyse naïve et les avantages de l'utilisation de la méthode décrite dans le présent article. Dans la simulation, les fichiers X et Y, contenant chacun 2 000 enregistrements, sont générés indépendamment 400 fois, chaque fichier est généré étant désigné par X(r) et Y(r), respectivement, avec r = 1, ..., 400. En particulier, sur X(r), x_i est tiré aléatoirement de la loi de Bernoulli de paramètre 0,5. Sur Y(r), y_i est tiré aléatoirement de la loi de Bernoulli de paramètre 0,5. Le r^e ensemble de données imparfaitement appariées, d*(r), est généré en appariant correctement chaque enregistrement du fichier Y(r) à un enregistrement du fichier X(r) avec la probabilité p = 0,8, 0,90, 0,95 et 1. Pour chaque r^e ensemble de données appariées, un échantillon pour examen manuel de 300 appariements est sélectionné. Chaque appariement contenu dans l'échantillon pour examen manuel est classé comme étant correct ou incorrect. Nous résumons la performance de l'estimateur du MV décrit à la section 3.2, et de la méthode naïve, qui repose sur l'hypothèse qu'il n'existe aucune erreur d'appariement,

6. Discussion

L'appariement des données est une méthode appropriée quand des ensembles de données doivent être jumelés en vue de renforcer la dimension temporelle ou des aspects tels que la portée ou la profondeur des renseignements. L'appariement des données est utilisé de plus en plus fréquemment par les organismes statistiques partout dans le monde. Il est bien connu que des erreurs peuvent se produire durant l'appariement des fichiers, par exemple quand des méthodes probabilistes sont appliquées. Toutefois, peu d'études traitant de la façon de faire des inférences valides en présence de ce genre d'erreurs ont été publiées. Le présent article offre des conseils méthodologiques et pratiques en vue d'aider les analystes dans ce domaine.

Tableau 5		*Quand p = 1, les estimateurs naïf et MV sont identiques par définition.	
Taux de couverture à 95 %	Erreur quadratique moyenne	Naïve	
		β ₀	β ₁
0,8	0,9	0,024	0,11
0,9	0,95	0,010	0,038
0,95	1	0,0056	0,016
		0,0043*	0,011*
		0,35	0,05
		0,80	0,62
		0,93	0,88
		93,0	94,25
		94,5	96,25

La première sont présentées ici). La section 3.3 avec R = 40 répétitions. L'EQM de β est calculée par

$$EQM(\beta) = \frac{1}{400} \sum_{r=1}^{400} (\hat{\beta}_r - \beta)(\hat{\beta}_r - \beta')$$

ou β_r est l'estimation du MV de β d'après d*(r).

Le tableau 5 montre que l'approche naïve produit de mauvais taux de couverture, à cause de son biais important en présence d'erreurs d'appariement et, par conséquent, son EQM relativement élevée. Pour le MV-méthode 1, les taux de couverture sont très proches des taux nominaux. Les résultats montrent que, si le pourcentage d'appariements corrects passe de 100 % à 80 %, l'EQM pour l'estimation MV augmente d'un facteur 3 environ pour β₀ et β₁ (les taux de couverture et l'EQM des méthodes 1 et 2 d'estimation du MV étant fort semblables, seuls les résultats pour la première sont présentés ici).

était de 86,6 %. L'amélioration produite par le MVC n'était pas importante, ce qui indique que le mécanisme sous-jacent qui donne des enregistrements non appartés ne dépendait pas de l'âge ni du sexe. Les estimations du PMV (voir la section 4) n'ont pas non plus amélioré beaucoup la situation, ce qui indique que le modèle logistique décrit à la section 5.1.4 n'expliquait pas le mécanisme générant les enregistrements non appartés. Curieusement, l'estimation du MV en utilisant le fichier EF était de 81,8 %, c'est-à-dire l'estimation de loin la plus proche de l'estimation de 78,3 % sous la NO. Or, dans le cas du seuil EF, les appartements incorrects sont la principale source d'erreur, c'est-à-dire le type d'erreur d'appariement que corrige l'estimateur du MV. Par conséquent, la correction des erreurs dues à des appartements incorrects a été beaucoup plus fructueuse que celle des erreurs dues à des enregistrements non appartés.

Les erreurs types des estimations NO, natives et MV sont indiquées entre parenthèses dans le tableau 2a. Dans le cas de TF et de EF, les erreurs-types de l'estimation MV sont environ 25 % et 75 % plus grandes, respectivement, que les erreurs-types de l'estimation native correspondantes. En outre, les erreurs-types de l'estimation MV pour EF sont légèrement plus faibles que pour TF, ce qui signifie que les appartements supplémentaires effectués en appliquant le

MV et leurs erreurs-types sont très proches.

Quel que soit le seuil utilisé, les estimations MV des tableaux 2a, b et c sont systématiquement plus proches de l'estimation NO que l'estimation native correspondante. Par exemple, dans le tableau 2b, l'estimation MV pour le seuil TF est de 36,9 %, c'est-à-dire appréciablement plus proche de l'estimation NO de 37,9 % que l'estimation native de 33,3 %. En fonction des estimations du tableau 2, nous pourrions soutenir que la décision d'utiliser le seuil TF ou EF n'est pas tellement importante, à condition d'utiliser l'estimateur du MV.

Le tableau 3 ressemble au tableau 2, à part le fait qu'il décrit les analyses des enregistrements appartés pour toutes les personnes de 15 ans et plus, plutôt que pour les Autochtones seulement. De nouveau, l'estimateur du MV produit systématiquement une amélioration dans le cas du seuil EF, mais non dans le cas du seuil TF. Le tableau 4 donne la situation d'étudiant en 2006 pour les personnes qui étaient étudiantes en 2005. Encore une fois, le MV donne généralement des estimations plus proches de l'estimation or correspondante, en particulier pour le seuil EF.

Tableau 2
Pourcentages d'Autochtones dans les diverses catégories d'emploi en 2006 sachant leur catégorie d'emploi en 2005. Pour chaque ensemble de données appartés, soit Seuil très faible ou extrêmement faible, les méthodes d'estimation peuvent être comparées à la norme d'or

Estimations pour divers ensembles de données appartés et méthodes									
a : Autochtones occupés en 2005									
Situation en 2006		or		Seuil très faible		MVC		Seuil extrêmement faible	
Occupé	78,3	86,7	86,0	86,4	86,6	86,1	71,9	81,8	
En chômage	(1,7)	(2,4)	(3,0)	4,1	4,1	4,2	(1,7)	(2,9)	
	3,7	4,2					6,3	3,3	
	(0,84)	(1,2)	(2,5)				(0,82)	(2,1)	
	17,8	9,0	9,3		9,1	9,6	21,6	14,7	
Inactif	(1,6)	(2,4)	(3,1)				(1,6)	(2,8)	
b : Autochtones en chômage en 2005									
Situation en 2006		or		Très faible		ML		Extrêmement faible	
Occupé	27,5	27,7	27,2	36,3	36,4	32,3	35,2	23,8	
En chômage	34,4	38,9	36,4				32,3	38,0	
	37,9	33,3					32,3	38,0	
c : Autochtones inactifs en 2005									
Situation en 2006		or		Naïve		ML		Naïve	
Occupé	13,7	10,8	10,7	7,4	7,4	6,3	24,3	10,5	
En chômage	5,8	7,6	81,5				6,3	5,8	
Inactif	80,4	81,5					69,2	83,5	

que nous disons très faible (TF), est considéré comme un seuil optimal, puisque, pour une gamme de seuils, les estimations naïves obtenues étaient les estimations « les plus proches » des estimations correspondantes produites en appliquant la NO (voir Bishop 2009). Le deuxième seuil, que nous disons extrêmement faible (EF), a effectivement pour objectif de maximiser le nombre d'enregistrements de la RGR qui sont apparés. Ci-après, nous désignons les deux fichiers d'enregistrements apparés sous la NB par les noms de leur seuil, TF et EF.

5.1.2 Résultats d'appariement

Sous la norme d'or (NO), 70 274 des 78 349 enregistrements de la RGR ont été apparés. Sous l'hypothèse que la NO correspond à l'appariement parfait, 8 075 personnes possédaient un enregistrement dans la RGR, mais non dans le recensement. En réalité, la NO n'est pas parfaite. Pour une discussion à ce sujet, voir Bishop, 2009.

Sous la norme de bronze avec seuil très faible (TF), 57 790 enregistrements de la RGR ont été apparés. Des 70 274 enregistrements de la RGR qui ont été apparés en appliquant la NO, 13 784 sont demeurés non apparés, 700 ont été apparés incorrectement et 55 790 ont été apparés correctement an appliquant le seuil TF. En outre, 1 300 enregistrements ont été apparés en appliquant le seuil TF, mais non en appliquant la NO et sont également des appartements incorrects. Donc, il y a eu, en tout, 2 000 (= 700 + 1 300) appartements incorrects.

Sous la norme de bronze avec seuil extrêmement faible (EF), 74 350 enregistrements de la RGR ont été apparés. Des 70 274 enregistrements de la RGR qui ont été apparés, en appliquant la NO, 2 811 sont demeurés non apparés, 9 793 ont été apparés incorrectement et 57 670 ont été apparés correctement en appliquant la norme de bronze avec le seuil EF. En outre, 6 887 enregistrement de la RGR ont été apparés sous la norme de bronze EF, mais non sous la NO.

En résumé, 97 % des appartements du fichier TF sont corrects et 20 % (=13 784/70 274) des enregistrements de la RGR apparés en appliquant la NO demeurent non apparés. Les chiffres correspondants pour le fichier EF sont de 78 % et 4 % (=2 811/70 274).

5.1.3 Modélisation de la probabilité qu'un appartement soit correct

Pour tous les appartements dans les conditions EF et TF, nous savons s'il s'agissait d'un appartement correct ou incorrect (par exemple, si un appartement EF est également produit par la NO, cet appartement est correct. Sinon, l'appariement EF est incorrect). Par conséquent, p_{xy} , de la section 3.1 était connue pour la NO. Toutefois, pour simuler la réalité, nous avons estimé p_{xy} , d'après un échantillon

5.1.4 Modélisation de la probabilité qu'un enregistrement demeure non apparié

À chaque enregistrement de la RGR apparié en appliquant la NO, nous avons attribué une variable indiquant si l'enregistrement était non apparié en appliquant la NB. Autrement dit, si l'enregistrement demeurait non apparié sous la NB, la variable indicatrice prenait la valeur « 1 » et autrement, la valeur « 0 ». Nous avons ajusté un modèle

logistique en utilisant la NO, où la variable dépendante était la variable indicatrice susmentionnée et les variables explicatives provenaient de la RGR. Le modèle contenait plus de 20 variables explicatives qui ont été choisies selon la méthode classique de sélection ascendante/descendante. Les variables explicatives comprennent le niveau de scolarité, la langue, la naissance outre-mer, le statut d'autochtone, et les indicateurs de variable clé manquante, tels que l'*ilot* (*meshblock*). La prédiction résultante a donné t_i qui est utilisé plus bas pour mettre en œuvre la méthode du pseudo-MV pour les tableaux de contingence ainsi que la régression logistique.

5.2 Résultats de l'analyse tabulaire

Le tableau 2 donne les résultats du recoupement des situations d'emploi des personnes autochtones déclarées à la RGR et au recensement. Le tableau 2a montre que en appliquant la NO, l'estimation de la proportion d'Autochtones occupés au recensement, sachant qu'ils étaient occupés au moment de la RGR, est de 78,3 %. L'estimation naïve correspondante dans les conditions TF, qui reposent sur l'hypothèse que les données sont parfaitement apparées, est de 86,7 %. Même après avoir remplacé chacun des 700 appartements TF incorrects par l'appariement correct correspondant et écarté les 1 300 enregistrements non apparés pour lesquels aucun appartement correct n'existe, l'estimation naïve demeure pratiquement inchangée, soit 86,0 % (*appariements or* dans le tableau 2a). nous voyons donc que la différence entre les estimations TF et NO ne résultent pas tellement d'appariements incorrects, et qu'elle est due principalement à des enregistrements non apparés. Cela explique partiellement pourquoi l'estimation du MV (86,4 %) dans les conditions EF (voir la section 3.1), qui comporte une correction pour les appartements incorrects seulement, ne produit qu'une faible amélioration. Le MV conditionnel (MVC) (voir la section 4) a été examiné pour essayer de réduire l'erreur due aux enregistrements non apparés risquant de donner lieu à une représentation incorrecte, pour ce qui est de l'âge et du sexe, dans le fichier de données apparées. L'estimation de l'emploi par le MVC

- la norme de bronze (NB) consistait à utiliser l'ilot et certains éléments de données du recensement (c'est-à-dire à ne pas se servir du nom et de l'adresse). Il s'agit d'une méthode qu'il est proposé d'utiliser pour les futurs travaux d'appariement de l'ABS.

Une description détaillée de l'étude de la qualité et de la méthode d'appariement figure dans Solon et Bishop (2009). Le rôle de la norme d'or (NO) dans l'étude de la qualité est essentiel. Elle fournit une référence par rapport à laquelle peut être évaluée la fiabilité de la norme de bronze (NB). L'utilité de la NO comme référence est due au fait que le nom et l'adresse sont des variables puissantes pour repérer les personnes figurant dans le recensement et dans la répétition générale du recensement et qu'elles ont été soumises à un examen manuel approfondi. Nous supposons donc que la NO correspond à l'appariement parfait. Par conséquent, les différences entre les estimations fondées sur la NO et sur la NB sont interprétées comme étant des erreurs. Autrement dit, nous nous intéressons à la fiabilité de la NB relativement à la NO.

5.1 Méthodologie d'appariement

5.1.1 Variables de groupe et d'appariement et algorithme d'affectation 1 – 1

À la présente sous-section, nous donnons un aperçu de l'appariement des enregistrements de la RGR à ceux du recensement en appliquant la norme de bronze (NB). La méthode d'appariement comprend une série de passages, où chacun est défini par un ensemble de variables de groupe et d'appariement, ainsi qu'un algorithme d'affectation 1-1. Dans le cas de passages multiples, seuls les enregistrements non appariés au premier passage peuvent être appariés au deuxième, seuls les enregistrements non appariés au deuxième passage peuvent être appariés au troisième, et ainsi de suite.

Le tableau 1 donne les variables de groupe, désignées par « G » pour la NB. Par exemple, durant le passage 1, un enregistrement du recensement et un enregistrement de la RGR ne sont considérés comme un appariement possible que s'ils contiennent la même valeur pour l'ilot.

Les variables d'appariement sont utilisées pour mesurer le degré de concordance entre une paire d'enregistrements. Un haut niveau de concordance donne à penser que la probabilité que la paire d'enregistrements constitue un appariement correct est élevée. Le tableau 1 donne les variables d'appariement, désignées par « A », pour la NB. Par exemple, durant le passage 1 pour la NB, une gamme de variables, dont le jour, le mois et l'année de la naissance, le pays de naissance et le plus haut niveau de qualification sont utilisés comme variables de couplage.

Tableau 1
Exemple de variables de groupe (G) et d'appariement (A) utilisées pour appairer les données du Recensement de 2006 avec celles de la répétition générale du recensement. Différentes variables de groupe ont été utilisées lors de chacun des deux passages

Variable	Passage 1	Passage 2
Jour de la naissance	A	G
Mois de la naissance	A	G
Année de la naissance	A	G
Sexe	A	G
Statut d'Autochtone	A	A
Pays de naissance	A	A
Langue parlée	A	A
Année de l'arrivée	A	A
État matrimonial	A	A
Religion	A	A
Domaine d'études	A	A
pour la qualification la plus élevée	A	A
Niveau de la qualification la plus élevée	A	A
Plus haut niveau de scolarité	A	A
Ilot (Mesh block)	G	A

À chaque passage, la sortie comprend un score pour chaque paire d'enregistrements. Le score est une mesure du degré de concordance entre les enregistrements formant la paire. Nous remettons à plus tard la définition formelle du score (pour des détails, voir (3.6), Conn et Bishop 2006), mais nous illustrons plus bas comment il peut être interprété. Considérons la NB au passage 2, pour lequel les enregistrements d'une paire contiennent la même date de naissance complète et le même sexe; une paire d'enregistrements recevra un score de 23,5 s'il y a concordance pour l'ilot (+17) et l'année d'arrivée (+8) et désaccord pour la religion (-1,5) (dans cet exemple, la situation de concordance pour les autres variables d'appariement contribuerait également au score, mais nous les ignorons pour simplifier l'illustration). La contribution de la concordance pour l'ilot (+17) est plus grande celle de la concordance concernant l'année d'arrivée (+8) parce que la première est moins susceptible que la seconde d'avoir lieu par chance uniquement.

Afin de formaliser la cible de l'algorithme d'appariement, désignons le score pour l'enregistrement i de la RGR et l'enregistrement j du recensement durant le passage p en appliquant la NB par r_{pij} . L'ensemble des scores des paires d'enregistrements r_{pij} et le seuil d'exclusion f_p sont utilisés par le progiciel d'appariement *Febrl* (voir Christen et Churches 2005) pour déterminer l'ensemble optimal d'appariements pour le passage p . Le terme f_p est la valeur minimale du score pour qu'une paire d'enregistrements soit considérée comme un appariement durant le passage p . L'algorithme *Febrl* cherche à maximiser $\sum_i r_{pij}$, sous la contrainte $r_{pij} > f_p$. De toute évidence, le nombre d'appariements dépend de f_p .

Dans la suite, nous évaluons la NB au moyen de deux ensembles différents de seuils, où un ensemble de seuils est défini par les seuils des passages 1 et 2. Le premier seuil,

A. Définissons $\mathbf{A} = (\mathbf{\Pi}_1, \dots, \mathbf{\Pi}_h, \dots, \mathbf{\Pi}_H)$, où $\mathbf{\Pi}_h = (\pi_{1h}^*, \dots, \pi_{gh}^*, \dots, \pi_{gh}^*, \dots, \pi_{gh}^*)$ et π_{gh}^* est la probabilité que $y_i = c$, $x_i = g$ et $\zeta_i = h$. L'estimateur du MV de $\mathbf{\Pi} = (\pi^{cx})$ de la section 2.1 quand les erreurs d'ap-
partement ne peuvent pas être ignorées est $\hat{\mathbf{\Pi}} = (\hat{\pi}^{cx})$, où

$$\hat{\pi}^{cx} = \sum_H \pi^{cxh} \hat{\pi}^{h|cx}, \quad (14)$$

où

$$\pi^{cxh} = \hat{\pi}^{cxh} \left(\sum_c \hat{\pi}^{c|xh} \right)^{-1}, \quad (15)$$

$\hat{\pi}^{cxh} = \sum_{i \in U_i} \hat{w}_{ic|xh}^*$, $\sum_{i \in U_i} \hat{w}_{ic|xh}^*$ est la somme sur les n^* enregistrements apprêtés et $\hat{\pi}^{h|cx}$ pour $h = 1, \dots, H$ est l'estimation classique de la distribution marginale de ζ sachant x dans le fichier X. En outre, si $i \notin s_c$,

$$\hat{w}_{ic|xh}^* = w_{ic|xh}^* \hat{p}^{xy^*h} + (1 - \hat{p}^{xy^*h}) \hat{\pi}^{c|xh}, \quad (16)$$

\hat{p}^{xy^*h} est la probabilité que le i^e appartement soit correct sachant que $x_i^* = x$, $\zeta_i^* = h$ et $y_i^* = y$, $w_{ic|xh}^* = 1$ si $x_i = x$, $\zeta_i = h$ et $y_i^* = y$, et $w_{ic|xh}^* = 0$ autrement. Si $i \in s_c$, alors $\hat{w}_{ic|xh}^* = w_{ic|xh}^*$, s'il est déterminé que l'appartement est correct et $\hat{w}_{ic|xh}^* = \hat{\pi}^{c|xh}$, s'il est déterminé qu'il est incorrect.

L'estimateur du MV $\hat{\pi}^{cx}$ s'obtient par itération entre (14), (15) et (16) jusqu'à la convergence.

4.3 Pseudo-maximum de vraisemblance (PMV)

À la présente section, nous discutons d'une alternative au MVC décrit à la section 4.2, que nous appelons pseudo-maximum de vraisemblance (voir Chambers et Skinner 2003). Il s'agit essentiellement d'une approche de pondération, qui pourrait être plus facile à mettre en œuvre que le MVC et qui s'appuie sur la factorisation donnée à la section 4.2. Elle comprend la résolution de versions pondérées des fonctions de score, Score(π ; \mathbf{d}) = $\mathbf{0}_{C-1}$ et Score(β ; \mathbf{d}) = $\mathbf{0}_K$ pour trouver π et β respectivement, où le poids d'un enregistrement demeure non apparié. Nous désignons la probabilité que l'enregistrement i ne restera pas non apparié par $t_i = E(\gamma_i)$, de sorte que les poids unitaires sont donnés par $q_i = t_i^{-1}$, où ici $i = 1, \dots, n^*$. Par conséquent, l'estimateur du PMV pour π^{cx} est

$$\hat{\pi}_{PMV}^{cx} = \hat{n}^{cx} \left(\sum_c \hat{n}^{c|x} \right)^{-1}, \quad (17)$$

où $\hat{n}^{c|xy} = \sum_{i \in U_i} q_i \hat{w}_{ic|x}^*$. L'estimation de $\hat{\pi}_{PMV}^{cx}$ s'obtient par itération entre la mise à jour de $\hat{w}_{ic|x}^*$, donné par (7), et (17)

5. Étude empirique

L'Australian Bureau of Statistics a réalisé une étude de qualité comportant l'appariement des enregistrements du Recensement de la population et du logement de 2006 aux enregistrements de la répétition générale du recensement (RGR). Durant la répétition générale du recensement, des renseignements ont été recueillis auprès de 78 349 personnes, un an avant le recensement. Durant le Recensement de 2006, des renseignements ont été recueillis auprès de plus de 19 millions de personnes. Pendant une brève période, durant laquelle les données du Recensement de 2006 ont été traitées, les noms et adresses étaient disponibles pour le recensement ainsi que pour la répétition générale du recensement. Durant cette période, les deux fichiers d'enregistrements au niveau de la personne ont été appariés en utilisant des normes d'information différentes :

- la norme *d'or* (NO) consistait à utiliser le nom, l'adresse, l'ilot (*mesh block*) et certains éléments de données du recensement. L'ilot est une zone géographique contenant habituellement 50 logements. Tous les noms et adresses ont été détruits à la fin de la période de traitement des données du recensement.

Pour illustrer la situation où les enregistrements non appariés ne peuvent pas être ignorés, considérons l'appariement d'une base de données contenant des renseignements personnels sur la situation d'emploi à une autre contenant des renseignements sur le niveau de scolarité. En outre, supposons que l'âge et le sexe, qui sont des variables corrélées à l'emploi et à la scolarité, sont disponibles dans l'une des bases de données. Après avoir effectué un examen manuel, nous pourrions constater que les enregistrements pour les jeunes hommes ont 50 % de chances de plus de rester non appariés que les enregistrements pour les femmes. Cela pourrait tenir au fait que les hommes sont moins susceptibles que les femmes de fournir des renseignements personnels qui sont utiles pour l'appariement. Clairement, dans le fichier de données appariées, il convient de donner aux enregistrements des hommes une pondération double de celle des enregistrements des femmes afin que l'analyse conjointe de la situation d'emploi et du niveau de scolarité soit sans biais.

cette variable ne figure pas dans le modèle logistique ou dans le tableau de contingence. À la section 4.2, nous expliquons comment le faire pour les tableaux de contingence. À la section 4.3, nous discutons d'une approche fondée sur la pseudo-vraisemblance qui consiste à attribuer aux enregistrés non apparus des pondérations visant à tenir compte de toute sur ou sous-représentation de certaines sous-populations dans les données apparues. De nouveau, le choix des pondérations devrait être justifié après l'analyse du sous-échantillon, s_{xc} , qui indique quels sont les enregistrés non apparus. Cet aspect est examiné plus en détail dans le contexte de l'étude empirique.

4.1 Pourquoi-nous ignorer les enregistrés non apparus ?

Définissons la variable $y_i = 1$ si l'enregistré i du fichier X est non apparié et $y_i = 0$ autrement. Soit aussi ζ_i une variable telle que $\zeta_i = 1, 2, \dots, h, \dots, H$, où H est le nombre de catégories pour ζ_i . Nous pouvons ignorer le fait que des enregistrés sont non apparus si nous sommes prêts à supposer que, conditionnellement à \mathbf{x}_i , les distributions de y_i , y_i et δ_i sont indépendantes. Techniquement, cette hypothèse mène à la factorisation suivante

$$p(y_i, \mathbf{x}_i, \delta_i, y_i, \zeta_i) \propto p(y_i | \mathbf{x}_i, \theta) p(\delta_i | \mathbf{x}_i) p(y_i | \mathbf{x}_i) p(\zeta_i)$$

où, de nouveau, $\theta = \beta$ ou Π . Il vaut la peine de vérifier si cette hypothèse est valide dans le cas du sous-échantillon pour examen manuel. Si l'hypothèse est raisonnable, il n'est pas nécessaire d'appliquer les méthodes décrites aux sections 4.2 et 4.3, et les méthodes de la section 3 suffisent.

Nous pourrions ne pas vouloir émettre l'hypothèse sus-mentionnée, mais être prêts à supposer que, conditionnellement à \mathbf{x} et ζ , les distributions de y_i , y_i et δ_i sont indépendantes. Dans ce cas, nous disons que les enregistrés non apparus sont ignorables. Techniquement, cette hypothèse mène à la factorisation suivante.

$$p(y_i, \mathbf{x}_i, \delta_i, y_i, \zeta_i) \propto p(y_i | \mathbf{x}_i, \zeta_i; \mathbf{A}) p(\delta_i | \mathbf{x}_i; \tau) p(y_i | \mathbf{x}_i, \zeta_i) p(\zeta_i)$$

où \mathbf{A} est le paramètre pour la distribution de $y_i | \mathbf{x}_i, \zeta_i$. Si nous nous intéressons à $p(y_i | \mathbf{x}_i; \theta)$, mais non à $p(y_i | \mathbf{x}_i, \zeta_i; \mathbf{A})$, une approche consiste à éliminer ζ_i de cette dernière par intégration (c'est-à-dire en prenant la moyenne sur toutes les valeurs possibles).

4.2 Maximum de vraisemblance conditionnel (MVC) pour les tableaux de contingence

Premièrement, paramétrisons la distribution conjointe de y_i , x_i et ζ_i par la distribution multinomiale de paramètre

La première méthode consiste à conditionner l'analyse sur une variable $\zeta_i = \zeta_i(z_i)$. La variable ζ est définie de façon qu'en présence d'enregistrements non apparus, l'inférence soit sans biais conditionnellement à ζ . Le terme ζ est introduit parce que, dans de nombreux cas, il serait peu commode ou inutile de conditionner sur toute l'information contenue dans \mathbf{z} . Il est possible de donner à ζ_i une valeur non manquante, même quand \mathbf{z}_i contient des valeurs manquantes. La forme exacte de la fonction $\zeta(\mathbf{z})$ devrait être justifiée après l'analyse du sous-échantillon, s_{xc} . Par exemple, si les personnes de moins de 20 ans sont sous-représentées dans le fichier de données apparues, ζ indiquerait si une personne a moins de 20 ans. Un moyen d'approcher l'analyse serait d'inclure ζ comme covariable dans le modèle de régression. La méthode décrite à la section 3 s'appliquerait alors directement. Cependant, les analystes pourraient souhaiter procéder à l'intégration sur ζ afin que

soit non apparus, soit apparus, soit non apparus à des enregistrés du fichier Y . Les enregistrés apparus du sous-échantillon doivent être catégorisés comme étant correctement ou incorrectement apparus durant le processus d'examen manuel. Un enregistré doit être classé comme étant *apparié* ou *autre*. *Non apparié* signifie que l'enregistré correspondait à être trouvé dans le fichier Y , mais qu'il n'y a pas eu d'appariement, tandis que *autre* indique que l'enregistré correspondait à pas été découvert dans le fichier Y et est par conséquent considéré comme non existant. Cette seconde classification pourrait être beaucoup plus difficile et plus longue que la première, parce qu'elle suppose qu'un autre processus exempt d'erreur existe pour vérifier qu'il est correct que certains appariements n'aient pas été faits. De par leur nature, les enregistrés non apparus contiennent peu d'information permettant de déterminer quel est l'appariement correct, même durant l'examen manuel. Ce genre de processus peut ne pas exister, auquel cas la correction pour tenir compte des enregistrés non apparus semble impossible. Cependant, il pourrait comprendre un examen manuel des noms qui figurent dans les deux fichiers à appairer. Par exemple, la personne chargée de l'examen manuel pourrait se rendre compte que les noms *John O. Smith* et *Joh O. Smith* qui figurent dans deux enregistrés distincts pourraient, en fait, faire référence à la même personne (avec un « n » manquant dans le deuxième cas, peut-être à cause d'erreur de lecture optique), tandis que le processus automatisé d'appariement pourrait traiter les deux noms comme étant entièrement différents. L'examinateur peut alors décider que les deux enregistrés susmentionnés correspondent à la même personne et devraient donc être appariés. Bishop (2009) et Wright (2009) discutent des avantages de l'examen manuel.

3.3 Estimation de la variance par le bootstrap

À la présente section, nous décrivons comment calculer la variance des estimations du MV de la section 3. Désignons le paramètre d'intérêt par θ , présenté plus haut, et l'estimation de son MV par $\hat{\theta}$. L'estimation par le bootstrap (Rubin et Little 2003) de la variance de $\hat{\theta}$, dénotée par $\hat{V}^{\text{boot}}(\hat{\theta})$, s'obtient comme il suit :

1. Tirer un échantillon répété de taille n_x du fichier de données appariées, \mathbf{d}^* , par échantillonnage aléatoire simple avec remise. Désigner le r^e échantillon répété par $\mathbf{d}^*(r)$. Le r^e échantillon répété est $s_c(r) = s_c \cap \mathbf{d}^*(r)$.
2. Calculer $\hat{\theta}(r)$ qui a la même forme que $\hat{\theta}$ excepté que $\mathbf{d}^*(r)$ est utilisé au lieu de \mathbf{d}^* et que $s_c(r)$ est utilisé au lieu de s_c .
3. Répéter les étapes 1 et 2 R fois, où R est le nombre de répétitions.
4. Calculer

$$\hat{V}^{\text{boot}}(\hat{\theta}) = \frac{1}{R} \sum_{b=1}^R (\hat{\theta}(b) - \bar{\hat{\theta}})(\hat{\theta}(b) - \bar{\hat{\theta}}).$$

4. Analyse en présence d'appariements incorrects et d'enregistrements non appariés

À la présente section, nous discutons de deux moyens d'analyser les données appariées en présence d'appariements incorrects et d'enregistrements non appariés. Comme nous l'avons mentionné dans l'introduction, le problème de l'analyse en présence d'enregistrements non appariés possède des similitudes évidentes avec le problème de la non-réponse d'une unité, ou non-réponse totale. L'existence d'enregistrements non appariés peut entraîner la sur ou la sous-représentation de certaines caractéristiques dans le fichier de données appariées, ce qui peut biaiser l'analyse. Comme nous l'exposons de manière plus détaillée plus bas, nous tirons parti du fait que le mécanisme donnant lieu aux enregistrements non appariés ne peut dépendre que de \mathbf{z} .

Nous considérons ici deux méthodes d'inférence en présence d'appariements incorrects et d'enregistrements non appariés, où les enregistrements appariés sont désignés au moyen de l'indice $i = 1, \dots, n^*$ (rappelons que le i^e enregistrement du fichier X est un enregistrement non apparié si $i \in U_{xy}$ et que l'enregistrement i n'a été apparié à aucun enregistrement du fichier Y). La méthode consiste à modéliser de façon indépendante les processus qui déterminent quels enregistrements sont appariés incorrectement et lesquels sont non appariés (voir la section 5 pour un exemple). Ces modèles requièrent qu'un sous-échantillon, désigné par s_{xc} , des enregistrements du fichier X soit soumis à un examen manuel. Les enregistrements compris dans le

\tilde{U}_i a la même forme que U_i excepté que \mathbf{b} est remplacé par $\hat{\mathbf{b}}_{xy^*}$, et \hat{p}_{xy^*} est la proportion estimée d'appariements corrects dans l'échantillon pour examen manuel pour chaque combinaison de \mathbf{x} et y^* .

où $r_{ik} = y_i^* x_{ik}$. L'espérance de r_{ik} conditionnellement à \mathbf{d}^* est

$$E^{d^*} (r_{ik} | \mathbf{x}_i = \mathbf{x}, y_i^* = y^*) = [y_i^* \hat{p}_{xy^*} + (1 - \hat{p}_{xy^*}) U_i] x_{ik} \quad \text{si } i \notin s_c$$

$$= y_i^* x_{ik} \quad \text{si } i \in s_c \text{ et } \delta_i = 1$$

$$= U_i x_{ik} \quad \text{si } i \in s_c \text{ et } \delta_i = 0$$

et \hat{p}_{xy^*} est la probabilité qu'un appariement avec $x_{ik} = 1$ soit correct sachant que $y_i^* = y^*$. L'estimateur du MV est alors obtenu par itération en vue de trouver la solution, désignée par $\hat{\mathbf{b}}$, pour \mathbf{b} dans (5) avec r_{ik} remplacé par \hat{r}_{ik} , où

$$\hat{r}_{ik} = [y_i^* \hat{p}_{xy^*} + (1 - \hat{p}_{xy^*}) \tilde{U}_i] x_{ik} \quad \text{si } i \notin s_c$$

$$= y_i^* x_{ik} \quad \text{si } i \in s_c \text{ et } \delta_i = 1$$

$$= \tilde{U}_i x_{ik} \quad \text{si } i \in s_c \text{ et } \delta_i = 0,$$

dans l'échantillon pour examen manuel pour chaque combinaison de \mathbf{x} et y^* . Autrement dit, si $y_i^* = 1$,

$$p_{ky^*} = \left(\sum_{i \in s_c} (1 - y_i^*) x_{ik} \delta_i \right) \left(\sum_{i \in s_c} (1 - y_i^*) x_{ik} \right)^{-1}$$

$$p_{ky^*} = \left(\sum_{i \in s_c} y_i^* x_{ik} \delta_i \right) \left(\sum_{i \in s_c} y_i^* x_{ik} \right)^{-1}$$

Cette approche ne nécessite le calcul que de $2K$ probabilités d'après l'échantillon pour examen manuel et, en ce sens, pourrait être préférable à l'approche décrite à la section 3.2.1 qui requiert le calcul d'un plus grand nombre de probabilités.

manuel convenant pour toutes les analyses possibles serait difficile.

Nous factorisons la distribution conjointe $p(y_i, x_i, \delta_i)$

$$p(y_i | x_i; \theta) p(x_i | \theta) p(\delta_i | x_i), \quad (6)$$

où $\theta = \beta$ dans le cas de la régression, et $\theta = \Pi$ dans le cas du tableau de contingence. La factorisation (6) signifie que les appariements sont incorrects au hasard (IAH), autrement dit, que les distributions $y_i | x_i$ et $\delta_i | x_i$ sont indépendantes. Sous cette hypothèse, il suffit de maximiser la vraisemblance associée au facteur $p(y_i | x_i; \theta)$. Tout au long de la présente section, nous faisons l'hypothèse (6). Il importe de souligner que (6) et le développement qui suit ne s'appuient sur aucune hypothèse nécessitant que le fichier X soit un sous-ensemble du fichier Y (par exemple quand les unités du fichier X sont un sous-échantillon des unités du fichier Y) ou que le processus d'appariement ne comporte qu'un seul passage. Nous supposons également que l'exactitude de l'appariement, δ_i , est indépendante d'un enregistrement à l'autre.

Comme nous l'avons mentionné dans l'introduction, un score est attribué à chaque enregistrement apparié en se basant sur la probabilité que les enregistrements appariés tiennent à la même unité. Désignons le score par r_i . Un examinateur a suggéré d'utiliser r_i pour paramétriser plus exactement la distribution de δ_i . Techniquement, cette suggestion nécessiterait le remplacement de $p(\delta_i | x_i)$ par $p(\delta_i | x_i, r_i)$ dans (6) et réduirait vraisemblablement la variabilité des estimateurs du MV discutés à la section 3. Il serait intéressant d'explorer cette piste dans le cadre de futurs travaux.

3.1 Tableaux de contingence

Soit $w_{ic|x}^* = 1$ si $y_i^* = c$ et $x_i = x$, et $w_{ic|x}^* = 0$ autrement. L'espérance de $w_{ic|x}^*$ sachant \mathbf{d}_i^* est

$$E^{d|x^*}(w_{ic|x}^* | x_i = x, y_i^* = y^*) =$$

$$w_{ic|x}^* p_{xy^*}^* + (1 - p_{xy^*}^*) \pi_{c|x}^* \quad \text{si } i \notin s_c$$

$$= w_{ic|x}^* \quad \text{si } i \in s_c \text{ et } \delta_i = 1$$

$$= \pi_{c|x}^* \quad \text{si } i \in s_c \text{ et } \delta_i = 0$$

et $p_{xy^*}^*$ est la probabilité que le i^e appariement soit correct sachant que $x_i = x$ et $y_i^* = y^*$. L'estimateur du MV de $\pi_{c|x}^*$ en se servant des données appariées de manière probabiliste, \mathbf{d}_i^* , est alors

$$\hat{\pi}_{c|x}^* = \hat{w}_{ic|x}^* \left(\sum_c \hat{w}_{ic|x}^* \right)^{-1} \quad (7)$$

$$\hat{w}_{ic|x}^* = \sum_i \tilde{w}_{ic|x}^*, \quad (8)$$

où

$$\tilde{w}_{ic|x}^* = w_{ic|x}^* \hat{p}_{xy^*}^* + (1 - \hat{p}_{xy^*}^*) \pi_{c|x}^* \quad \text{si } i \notin s_c$$

$$= w_{ic|x}^* \quad \text{si } i \in s_c$$

$$= \pi_{c|x}^* \quad \text{si } i \in s_c \text{ et } \delta_i = 0$$

$$\hat{p}_{xy^*}^* = \left(\sum_{i \in s_c} w_{ic|x}^* \delta_i \right) \left(\sum_{i \in s_c} w_{ic|x}^* \right)^{-1}. \quad (10)$$

La procédure d'estimation consiste à itérer (7), (8) et (9) jusqu'à la convergence. Plus précisément, l'algorithme est :

1. Calculer $\hat{p}_{xy^*}^*$ d'après (10).
2. Initialiser $\hat{\pi}_{c|x}^{(0)}$ puis calculer $\hat{w}_{ic|x}^{(0)}$ d'après (9) et ensuite $\hat{\pi}_{c|x}^{(0)}$ d'après (8).
3. Calculer $\hat{\pi}_{c|x}^{(i)}$ d'après (7) en utilisant $\hat{\pi}_{c|x}^{(i-1)}$.
4. Calculer $\hat{w}_{ic|x}^{(i)}$ d'après (9) en utilisant $\hat{\pi}_{c|x}^{(i)}$ puis calculer $\hat{\pi}_{c|x}^{(i)}$ d'après (8) en utilisant $\hat{w}_{ic|x}^{(i)}$.
5. Itérer 3 et 4 jusqu'à la convergence.

La valeur initiale $\hat{\pi}_{c|x}^{(0)}$ pourrait être choisie comme étant l'estimation naïve de $\pi_{c|x}^*$ décrite plus haut à la section 3. Cependant, nous avons constaté que le choix de la valeur initiale n'est pas important.

3.2 Régression logistique

Nous décrivons ci-après deux méthodes du MV (méthodes 1 et 2) pour estimer β en se servant de données appariées de manière probabiliste, \mathbf{d}_i^* . Les deux méthodes donnent des estimations sans biais sous l'hypothèse que les appariements sont incorrects au hasard. Elles se distinguent par le niveau d'aggrégation auquel sont estimées les probabilités qu'un appariement soit correct. La méthode 1 requiert l'obtention de ces probabilités à un niveau plus fin d'aggrégation, ce qui pourrait signifier qu'elle produit des estimations plus variables que la méthode 2.

3.2.1 Méthode 1

L'espérance de y conditionnellement aux données appariées est

$$E^{d|x^*}(y_i | x_i = x, y_i^* = y^*) =$$

$$y_i^* p_{xy^*}^* + (1 - p_{xy^*}^*) v_i \quad \text{si } i \notin s_c$$

$$= y_i^* \quad \text{si } i \in s_c \text{ et } \delta_i = 1$$

$$= v_i \quad \text{si } i \in s_c \text{ et } \delta_i = 0$$

et $p_{xy^*}^*$ est la probabilité que le i^e appariement soit correct sachant que $x = x_i$ et $y^* = y_i^*$.

L'estimateur du MV est alors obtenu par itération en vue de trouver la solution, désignée par $\hat{\beta}$, pour β dans (5) avec y_i remplacé par \tilde{y}_i , où

3. Analyse en présence d'appariements incorrects

À la présente section, nous considérons la situation où le fichier d'enregistrements appariés contient des appariements incorrects, mais aucun enregistrement non apparié. Cette situation se produit quand chacun des enregistrements du fichier X est apparié à un enregistrement du fichier Y (d'où $n_x \leq n_y$). Définissons le fichier d'enregistrements appariés par $\mathbf{d}^* = \{\mathbf{d}_i^* = (y_i^*, x_i^*) : i = 1, \dots, n_x\}$, où y_i^* est la valeur de y qui est *appariée* à l'enregistrement i du fichier X. Pour clarifier, y_i^* est la valeur vraie de y pour l'enregistrement i dans le fichier X, de sorte que $y_i^* = y_i$ si l'enregistrement i est apparié correctement.

L'estimateur donné par (2), associé à l'hypothèse que $y_i^* = y_i$ pour $i = 1, \dots, n_x$, est naïf, puisqu'il traite le fichier d'enregistrements appariés de manière probabiliste comme s'ils étaient parfaitement appariés. En général, nous devons les estimateurs du MV qui tiennent compte du fait que les données ont été appariées de manière probabiliste ou appariées imparfaitement d'une certaine façon.

Il est courant, en pratique, de tirer du fichier d'enregistrements appariés un sous-échantillon, désigné par s_c , puis de l'examiner manuellement. Durant l'examen manuel, un appariement, \mathbf{d}_i , est classé comme étant correct ou incorrect. Soit $\delta_i = 1$ si l'enregistrement i dans le fichier X est correctement apparié et $\delta_i = 0$ autrement.

La conception du sous-échantillon pour examen manuel est un problème important, surtout parce que l'examen manuel est souvent coûteux. Les utilisations possibles d'un échantillon pour examen manuel comprennent l'estimation de la proportion d'enregistrements correctement appariés et d'enregistrements non appariés, pour pouvoir décider quels enregistrements devraient être appariés et lesquels devraient rester non appariés, pour s'assurer que l'inférence en se servant de \mathbf{d}^* est correcte (c'est-à-dire l'objectif du présent article) ou pour déterminer comment pourraient être améliorée la façon dont les enregistrements sont appariés. (Dans les applications de l'ABS susmentionnées, les échantillons pour examen manuel ont été conçus de manière à s'assurer que chaque appariement ait au moins une probabilité spécifiée d'être correct.) Si l'on veut s'assurer que l'inférence en se servant de \mathbf{d}^* sera correcte, la sélection de l'échantillon pour examen manuel par échantillonnage aléatoire simple est une approche raisonnable. Un sous-échantillon pour examen manuel plus efficace pourrait peut-être être conçu, mais il n'existe aucun moyen évident de le faire, parce que les paramètres que nous devons estimer pour appliquer la méthode du MV décrite dans le présent article dépend de l'analyse particulière (par exemple choix de y et \mathbf{x}). Concevoir un échantillon pour examen

appariement parfait. Un appariement parfait signifie que chacun des enregistrements du fichier X est correctement apparié à l'enregistrement correspondant du fichier Y (c'est-à-dire qu'il n'y a pas d'appariement incorrect ni d'enregistrement non apparié). Sous appariement parfait, $n_y = n_x$ et l'ensemble d'enregistrements appariés est désigné par $\mathbf{d} = \{(y_i, x_i) : i = 1, \dots, n_x\}$. Sous appariement parfait, la fonction de score pour $\boldsymbol{\pi}^x = (\pi_{1^x}, \dots, \pi_{C^x})$ caractérisée par la distribution multinomiale est

$$\text{Score}(\boldsymbol{\pi}^x; \mathbf{d}) = (\text{Score}(\pi_{1^x}; \mathbf{d}), \dots, \text{Score}(\pi_{C^x}; \mathbf{d}))' \quad (1)$$

$$\text{Score}(\pi_{c^x}; \mathbf{d}) = \sum_i (w_{ic^x} \pi_{ic^x}^{-1} - w_{ic^x} \pi_{ic^x})$$

$$= n_{c^x} \pi_{c^x}^{-1} - n_{c^x} \pi_{c^x}$$

pour $c = 1, \dots, C - 1$, où $n_{c^x} = \sum_i w_{ic^x}$, $w_{ic^x} = 1$ si $y_i = c$ et $x_i = x$ et $w_{ic^x} = 0$ autrement, et la catégorie correspondant à $y = C$ est choisie arbitrairement comme catégorie de référence. La résolution de Score($\boldsymbol{\pi}^x; \mathbf{d}$) = $\mathbf{0}^{C-1}$ pour trouver $\boldsymbol{\pi}^x$, où $\mathbf{0}^{C-1}$ est un vecteur colonne de zéros de dimension $C - 1$, donne l'estimateur du maximum de vraisemblance (MV)

$$\hat{\pi}_{c^x} = n_{c^x} / n_x \quad (2)$$

où

$$n_x = \sum_i w_{ic^x}$$

$$\hat{\pi}_{c^x} = 1 - \sum_{c=1}^{C-1} \hat{\pi}_{c^x}$$

2.2 Régression logistique

Considérons le modèle de régression logistique

$$E(y_i) = v_i \quad (3)$$

$$v_i = 1 / [1 + \exp(\boldsymbol{\beta}' \mathbf{x}_i)] \quad (4)$$

Pour (4), les K éléments de \mathbf{x}_i sont des variables dichotomiques et y_i est maintenant une variable dichotomique disponible dans le fichier Y. Si nous définissons $\mathbf{x} = (x_1, \dots, x_{n_x}, x_{n_x+1}, \dots, x_{n_y})'$ et $\mathbf{v} = (v_1, \dots, v_{n_y})'$, la matrice de score pour $\boldsymbol{\beta}$ basée sur des données partiellement appariées, \mathbf{d} , est donnée par

$$\text{Score}(\boldsymbol{\beta}; \mathbf{d}) = \mathbf{x}' (\mathbf{y} - \mathbf{v}) \quad (5)$$

La résolution de Score($\boldsymbol{\beta}; \mathbf{d}$) = $\mathbf{0}^K$ pour trouver $\boldsymbol{\beta}$ donne l'estimation du MV, $\hat{\boldsymbol{\beta}}$, que l'on peut obtenir en appliquant la méthode bien connue de Newton-Raphson.

l'appariement correct. De manière plus générale, des cas d'enregistrements non appariés peuvent avoir lieu quand certaines sous-populations sont relativement difficiles à appairer. Par exemple, des champs tels que l'état matrimonial, la qualification, le domaine d'études et le plus haut niveau de scolarité ne seraient en général pas aussi puissants lorsque l'appariement concerne des enfants que lorsqu'il s'agit d'adultes ayant atteint la maturité. Dans cette situation, l'appariement des données doit décider s'il doit ou non appartenir ce genre d'enregistrements. Nous définissons l'ensemble d'enregistrements appariés par U_l de taille n^* de sorte que $n^* \leq n_x$ et $n^* \leq n_y$.

Le problème que pose l'analyse quand des enregistrements sont non appariés présente des points communs manifestes avec le problème de la non-réponse totale d'une unité. Dans les deux cas, un seul sous-ensemble d'enregistrements légitimes est disponible pour l'analyse. Le mécanisme de non-réponse dans les enquêtes par sondage dépend, en réalité, d'un ensemble inconnu de variables. En revanche, ici, nous avons le léger avantage de savoir que la probabilité qu'un enregistrement demeure non apparié ne peut être qu'une fonction de \mathbf{z} . Le problème de la non-réponse est souvent résolu par pondération ou au moyen d'un certain argument de conditionnement. Dans le présent article, nous envisageons les deux approches pour résoudre le problème des enregistrements non appariés.

Il existe un compromis naturel entre le nombre d'enregistrements non appariés et le nombre d'appariements incorrects (et par conséquent le biais qu'ils introduisent). Considérons le cas où le fichier X est un sous-échantillon du fichier Y de sorte que $U_{xy} = U_x$. L'appariement de tous les enregistrements du fichier X ne donnera lieu, par définition, à aucun enregistrement non apparié, mais maximisera le nombre d'appariements incorrects. Si nous décidons plutôt de ne former que des appariements dont nous sommes convaincus qu'ils sont corrects, le nombre d'appariements incorrects diminuera, mais le nombre d'enregistrements non appariés augmentera. En pratique, trouver l'équilibre optimal entre les biais dus aux enregistrements non appariés et aux appariements incorrects dépend de l'analyse qui doit être effectuée, de la méthode d'appariement et de leur interaction. Pour une discussion pratique approfondie de cette question, voir Bishop (2009).

Il convient de mentionner que le problème de l'inférence en présence d'appariements incorrects d'enregistrements est semblable au problème de l'inférence en présence de classifications incorrectes de la variable dépendante, qui est une forme d'erreur de mesure (voir Fuller 1987). Dans ce dernier cas, des hypothèses d'identification font la distinction entre le mécanisme d'erreur de classification et les mécanismes du modèle, et sont nécessaires puisqu'habituellement, on ne dispose d'aucune mesure exempte d'erreur.

2. Appariement parfait

À la section 2, nous résumons l'approche du MV pour les tableaux de contingence et l'analyse de régression sous appariement parfait. À la section 3, nous considérons l'appariement parfait du MV en présence d'appariements incorrects. À la section 4, nous examinons l'approche du MV en présence d'appariements incorrects et d'enregistrements non appariés. À la section 5, au moyen d'une étude empirique, nous démontrons l'efficacité de bon nombre des estimateurs proposés. Enfin à la section 6, nous résumons les résultats.

2.1 Tableaux de contingence

En ce qui concerne la notation, il est commode, lorsque l'on envisage l'analyse d'un tableau de contingence, de transformer \mathbf{x}_i en une variable catégorique unique x de sorte que $x = 1, 2, \dots, g, \dots, G$. Soit y une variable catégorique du fichier Y , où $y = 1, \dots, c, \dots, C$. Considérons la factorisation qui suit de la distribution de x et de y

$$p(y, x) = p_1(y) p_2(x),$$

où $\boldsymbol{\Pi} = (\boldsymbol{\pi}_1^y, \dots, \boldsymbol{\pi}_g^y, \dots, \boldsymbol{\pi}_G^y)'$, $\boldsymbol{\pi}_g^y = (\pi_{1|g}, \dots, \pi_{c|g}, \dots, \pi_{C|g})'$, $\pi_{c|g}$ est la probabilité que $y = c$ sachant $x = g$. Nous supposons que, pour chaque valeur de x , il existe C valeurs possibles de y , ce qui implique que la dimension de $\boldsymbol{\Pi}$ est CG .

Considérons maintenant l'estimation du maximum de vraisemblance du paramètre $\boldsymbol{\Pi}$, caractérisant p_1 , sous

fichiers à apparier soit un sous-ensemble de l'autre fichier. Fréquente en pratique, cette situation s'est présentée dans tous les exemples susmentionnés d'appariements à l'ABS. Il convient aussi de mentionner que les fichiers à apparier ne doivent pas nécessairement être reliés par un mécanisme d'échantillonnage, comme dans le cas où le fichier le plus petit est un sous-échantillon aléatoire des unités comprises dans le plus grand fichier. L'élimination de cette restriction signifie que les deux fichiers peuvent être des ensembles de données administratives.

Considérons l'appariement de deux fichiers désignés par X et Y . Le fichier Y contient la variable y de la population d'individus U_y comprenant n_y enregistrés. Le fichier X contient un vecteur, x , de variables sur la population d'individus U_x comprenant n_x enregistrés. L'inférence a pour cible la population de n_{xy} individus, désignée par $U_{xy} = U_x \cap U_y$, qui sont communs au fichier X et au fichier Y . Les fichiers X et Y contiennent également un vecteur de champs, désigné par z , qui sont utilisés pour apparier les fichiers en utilisant un algorithme d'appariement probabiliste. Naturellement, puisque nous considérons un appariement probabiliste ici, la variable z ne constitue pas un identificateur d'unité unique.

L'appariement des fichiers X et Y permet d'analyser la distribution conjointe de x et de y . Deux sources d'erreur peuvent avoir une incidence sur l'analyse de la distribution conjointe en se servant du fichier de données appariées. Ces erreurs correspondent aux *appariements incorrects* et aux *enregistrements non appariés*.

Un appariement est correct si les deux enregistrements appariés appartiennent à une même personne. Un appariement est incorrect si les deux enregistrements appariés n'appartiennent pas à la même personne. Les appariements incorrects peuvent accroître ou réduire artificiellement la corrélation entre x et y . Un exemple du second cas est l'appariement aléatoire, dans lequel les enregistrements du fichier X sont appariés aléatoirement aux enregistrements du fichier Y .

Le i^{e} enregistrement du fichier X est défini comme un *enregistrement non apparié* si $i \in U_{xy}$ et que l'enregistrement i n'a pas été apparié à un enregistrement du fichier Y . Autrement dit, un enregistrement non apparié est un enregistrement du fichier X qui pourrait être apparié correctement, mais n'a pas été apparié du tout (tout au long de l'exposé, nous adoptons la convention de définir les enregistrements non appariés en fonction du fichier X , mais la définition pourrait également être formulée en fonction des enregistrements du fichier Y). Il se pourrait que l'on ne puisse pas toujours apparier un enregistrement particulier du fichier X en ayant la certitude que l'appariement est correct. Cette situation peut se présenter s'il manque dans un enregistrement des champs qui sont utiles pour établir

2001), le National Center for Health Statistics des États-Unis (National Center for Health Statistics 2009) et l'Office fédéral de la statistique de la Suisse dans le cadre de son étude longitudinale des personnes vivant en Suisse.

L'appariement des données rend possibles de nouveaux produits et analyses statistiques. En général, traiter naïvement le fichier de données appariées de manière probabiliste comme s'il s'agissait d'un appariement parfait produit des estimations biaisées. Lahiri et Larsen (2005), ainsi que Scheuren et Winkler (1993) ont proposé des méthodes permettant d'estimer sans biais les coefficients d'un modèle de régression linéaire sous appariement probabiliste d'enregistrements. Plus récemment, Chambers et coll. (2009) et Chambers (2008) ont élargi la portée de ces travaux à un vaste ensemble de modèles en utilisant des équations d'estimation généralisées et, dans le cas de l'appariement de deux fichiers, en permettant que l'un des fichiers soit un sous-ensemble de l'autre.

Le présent article décrit l'élaboration d'une approche d'estimation du maximum de vraisemblance (MV) pour analyser les données appariées de manière probabiliste. La technique d'estimation est simple et mise en œuvre en se servant de l'algorithme EM bien connu. L'approche consistant à remplacer les statistiques qui seraient observées dans le cas de données parfaitement appariées par leur espérance conditionnellement aux données appariées. En supposant que l'espérance est spécifiée correctement, cette approche permet d'éviter les deux limites des travaux antérieurs décrites ci-après.

Premièrement, alors que les méthodes antérieures reposaient sur un appariement exécuté en un seul passage, l'appariement probabiliste comporte habituellement des passages multiples. Dans ce cas, seuls les enregistrements non appariés au moment du premier passage peuvent faire l'objet d'un appariement durant le deuxième passage, puis seuls les enregistrements non appariés au cours des deux premiers passages peuvent être appariés lors du troisième, et ainsi de suite. Chaque passage est conçu en vue d'apparier les enregistrements présentant un ensemble commun particulier de caractéristiques. Par exemple, le premier passage peut être conçu de manière à apparier les enregistrements appartenant aux personnes qui n'ont pas changé d'adresse entre les dates de référence des deux fichiers. Le deuxième passage peut être conçu pour tenir compte des changements d'adresse. Un exemple de ce genre d'approche est donné au tableau 1 à la section 5.

Deuxièmement, les méthodes antérieures reposaient sur l'hypothèse que les deux fichiers contenaient des enregistrements appartenant exactement aux mêmes unités ou que l'ensemble d'unités figurant dans un fichier était un sous-ensemble de celles présentes dans l'autre fichier. L'appariement que nous proposons ne requiert pas que l'un des

Estimation du maximum de vraisemblance pour les tableaux de contingence et la régression logistique en présence de données incorrectement appariées

James O. Chipperfield, Glenys R. Bishop et Paul Campbell¹

Résumé

L'appariement des données consiste à jumeler des enregistrements issus de deux fichiers ou plus que l'on pense appartenir à une même unité (par exemple une personne ou une entreprise). Il s'agit d'un moyen très courant de renforcer la dimension temporelle ou des aspects tels que la portée ou la profondeur des détails. Souvent, le processus d'appariement des données n'est pas exempt d'erreur et peut aboutir à la formation d'une paire d'enregistrements qui n'appartiennent pas à la même unité. Alors que le nombre d'applications d'appariement d'enregistrements croît exponentiellement, peu de travaux ont porté sur la qualité des analyses effectuées en se servant des fichiers de données ainsi appariées. Traiter naïvement ces fichiers comme s'ils ne contenaient pas d'erreurs mène, en général, à des estimations biaisées. Le présent article décrit l'élaboration d'un estimateur du maximum de vraisemblance pour les tableaux de contingence et la régression logistique en présence de données incorrectement appariées. Simple, cette méthode d'estimation est appliquée en utilisant l'algorithme EM bien connu. Dans le contexte qui nous occupe, l'appariement probabiliste des données est une méthode reconnue. Le présent article démontre l'efficacité des estimateurs proposés au moyen d'une étude empirique s'appuyant sur cet appariement probabiliste.

Mots clés : Appariement des données ; appariement probabiliste ; maximum de vraisemblance ; tableaux de contingence ; régression logistique.

1. Introduction

L'appariement des données, également appelé appariement ou couplage d'enregistrements, est la tâche consistant à jumeler des enregistrements que l'on pense appartenir à une même unité (par exemple une personne ou une entreprise) et qui sont tirés de deux fichiers ou plus. L'appariement des données est une technique indiquée pour jumeler des ensembles de données en vue de renforcer la dimension temporelle, ou des aspects tels que la portée ou la profondeur des détails. Dans des conditions idéales, l'appariement serait parfait, autrement dit seuls les enregistrements appartenant à la même unité seraient appariés et tous les appariements possibles seraient faits. Malheureusement, très souvent, il n'en est pas ainsi, surtout si l'on se sert pour appairer les enregistrements de champs pouvant contenir des valeurs incorrectes, des valeurs manquantes ou des valeurs légitimement différentes pour une unité particulière. On recourt souvent à l'appariement probabiliste quand les fichiers contiennent un ensemble commun de variables ou de champs qui fournissent des renseignements d'identification partiels, mais ne constituent pas un identificateur d'unité unique. Dans l'appariement probabiliste (Fellegeri et Sunter 1969), un score est attribué à chacun des appariements possibles en se basant sur la probabilité que les enregistrements appartiennent à la même unité. Ce score est calculé en comparant les valeurs des variables d'appariement qui sont communes aux deux fichiers. Un appariement

est alors déclaré si le score d'appariement est supérieur à un seuil donné. Un algorithme d'optimisation peut être utilisé pour s'assurer que chaque enregistrement d'un fichier ne soit pas apparié à plus d'un enregistrement d'un autre fichier. Les méthodes probabilistes d'appariement de fichiers sont bien établies aujourd'hui (voir Herzog, Scheuren et Winkler 2007, Winkler 2001 et Winkler 2005), et il existe toute une gamme de logiciels pour les mettre en œuvre.

Cette situation découle de l'importance continue de l'appariement dans divers domaines, particulièrement ceux touchant aux politiques en matière de santé et aux politiques sociales. Les exemples récents d'appariements probabilistes de données effectués par l'Australian Bureau of Statistics (ABS) comprennent l'appariement d'enregistrements provenant du Recensement de la population et du logement de l'Australie de 2006 avec ceux d'un certain nombre d'ensembles de données, dont les enregistrements des décès survenus en Australie (Australian Bureau of Statistics 2008), le Census Dress Rehearsal (Répétition générale du recensement) de 2006 (Solon et Bishop 2009) et l'Australian Migrants Settlements Database (Wright, Bishop et Ayre 2009). Dans le secteur australien de la santé, les méthodes d'appariement probabiliste sont employées par la Western Australian Data Linkage Unit (Holman, Bass, Rouse et Hobbs 1999) et par le New South Wales Centre for Health Record Linkage. Sur la scène internationale, les méthodes probabilistes sont utilisées par Statistique Canada (Fair 2004), le Census Bureau des États-Unis (voir Winkler

- Brick, Flores Cervantes, Lee et Norman : Erreurs non dues à l'échantillonnage dans les enquêtes téléphoniques
- Skinner, C.J., et Rao, J.N.K. (1996). Estimation in dual frame surveys with complex designs. *Journal of the American Statistical Association*, 91, 349-356.
- Steeh, C. (2004). A New Era for Telephone Surveys. Présenté à Annual Conference of the American Association for Public Opinion Research, Phoenix, AZ.
- Vicente, P., et Reis, E. (2009). The mobile-only population in Portugal and its impact in a dual frame telephone survey. *Survey Research Methods*, 3, 105-111.
- Tucker, C., Brick, J.M. et Meekins, B. (2007). Household telephone service and usage patterns in the U.S. in 2004: Implications for telephone samples. *Public Opinion Quarterly*, 71, 3-22.

Brick, J.M., Dipko, S., Presser, S., Tucker, C. et Yuan, Y. (2006). Nonresponse bias in a dual frame sample of cell and landline numbers. *Public Opinion Quarterly*, 70, 780-793.

Brick, J.M., Edwards, W.S. et Lee, S. (2007). Sampling telephone numbers and adults, interview length, and weighting in the California Health Interview Survey cell phone pilot study. *Public Opinion Quarterly*, 71, 793-813.

California Health Interview Survey (2009). CHIS 2007 Methodology Series: Rapport 4 – Response Rates. Los Angeles, CA : UCLA Center for Health Policy Research. Disponible au www.chis.ucla.edu/pdf/CHIS2007_method4.pdf.

Edwards, W.S., Brick, J.M. et Grant, D. (2008). Relative Costs of a Multi-frame, Multi-mode Enhancement to an RDD Survey. Présenté à Annual Conference of the American Association for Public Opinion Research, la Nouvelle-Orléans, LA.

Fleeman, A. (2007). Survey Research Using Cell Phone Sample: Important Operational and Methodological Considerations. Présenté à Annual Conference of the American Association for Public Opinion Research, Anaheim, CA.

Fuller, W.A., et Burmeister, L.F. (1972). Estimators for samples selected from two overlapping frames. *Proceedings of the Social Statistics Section*, 245-249.

Hartley, H.O. (1962). Multiple Frame Surveys. *ASA Proceedings of the Social Statistics Section*, 203-206.

Hartley, H.O. (1974). Multiple frame methodology and selected applications. *Sankhyā*, C, 36, 99-118.

Kalton, G., et Anderson, D.W. (1986). Sampling Rare Populations. *Journal of the Royal Statistical Society*, A 149, 65-82.

Kish, L. (1965). *Survey Sampling*. New York : John Wiley & Sons, Inc.

Kennedy, C. (2007). Evaluating the effects of screening for telephone service in dual frame RDD Surveys. *Public Opinion Quarterly*, 71, 750-771.

Keeter, S., Dimock, M. et Christian, L. (2008). Calling Cell Phones in '08 Pre-election Polls. News Release from The Pew Research Center for the People & the Press. Disponible au www.pewresearch.org/pubs/1061/cell-phones-election-polling.

Kuusela, V., Calleagar, M. et Vehovar, V. (2008) The influence of mobile telephones on telephone surveys. Dans *Advances in Telephone Survey Methodology*, (Eds., J.M. Lepkowski, C. Tucker, J.M. Brick, E.D. de Leeuw, L. Japel, P.J. Lavrakas, M.W. Link et R.L. Sangster), New York : John Wiley & Sons, Inc., Chapitre 4, 87-112.

Lohr, S. (2009). Multiple frame surveys. Dans *Handbook of Statistics: Sample Surveys Design, Methods and Applications*, (Ed., D. Pfeffermann), Elsevier, Amsterdam, Chapitre 4, Vol. 29A.

Lohr, S., et Rao, J.N.K. (2000). Inference in dual frame surveys. *Journal of the American Statistical Association*, 95, 271-280.

Lohr, S., et Rao, J.N.K. (2006). Estimation in multiple-frame surveys. *Journal of the American Statistical Association*, 101, 1019-1030.

Skinner, C.J. (1991). On the efficiency of taking ratio estimation for multiple frame surveys. *Journal of the American Statistical Association*, 86, 779-784.

PC2. De tous les appels téléphoniques que vous recevez, est-ce que ...

[Altermiez les options en gardant CERTAINS au milieu.]

Tous ou presque tous sont reçus sur un téléphone mobile ?

Certains sont reçus sur un téléphone mobile et certains, sur un téléphone ordinaire à la maison ?

Tous ou presque tous sont reçus sur un téléphone ordinaire à la maison ?

Pew Research Center for the People & The Press –

Téléphone fixe

[TRADUCTION]

PL1. Maintenant, en pensant à votre utilisation du téléphone... Possédez-vous un téléphone mobile qui fonctionne ?

[Si la réponse est oui, posez PL2.]

PL2. De tous les appels téléphoniques que vous recevez, est-ce que ...

[Altermiez les options en gardant CERTAINS au milieu]

Tous ou presque tous sont reçus sur un téléphone mobile ?

Certains sont reçus sur un téléphone mobile et certains, sur un téléphone ordinaire à la maison ?

Tous ou presque tous sont reçus sur un téléphone ordinaire à la maison ?

Bibliographie

Bankier, M.D. (1986). Estimators based on several stratified samples with applications to multiple frame surveys. *Journal of the American Statistical Association*, 81, 1074-1079.

Blumberg, S.J., et Luke, J.V. (2009). Wireless Substitution: Early Release of Estimates from the National Health Interview Survey, juillet-décembre 2008. National Center for Health Statistics. Disponible au <http://www.cdc.gov/nchs/nhis.htm>.

Blumberg, S.J., Luke, J.V., Davidson, G., Davern, M.E., Yu, T. et Soderberg, K. (2009). Wireless substitution: State-level estimates from the National Health Interview Survey, janvier-décembre 2007. Hyattsville, MD : National Center for Health Statistics. *National Health Statistics Reports*, 14.

Brick, J.M., Brick, P.D., Dipko, S., Presser, S., Tucker, C. et Yuan, Y. (2007). Cell phone survey feasibility in the U.S.: Sampling and calling cell numbers versus landline numbers. *Public Opinion Quarterly*, 71, 29-33.

dans les enquêtes à base de sondage double, il est important de prendre en considération la façon dont le choix de λ affecte le biais et l'erreur quadratique moyenne des estimations. Les autres méthodes d'échantillonnage et d'estimation discutées dans le présent article peuvent également s'appliquer à d'autres enquêtes à base de sondage double. L'utilité de ces méthodes dépend de la compréhension de la nature des erreurs non dues à l'échantillonnage ainsi que de la disponibilité des données auxiliaires qui pourraient être utilisées dans le calage.

Remerciements

Nous remercions Scott Keeter, Stephen Blumberg et Julian Luke d'avoir fourni les données sur lesquelles s'appuie le présent article. Nous remercions également de nombreuses personnes de leurs commentaires constructifs concernant des versions antérieures du manuscrit, y compris Sherm Edwards, Ralph DiGaetano, David Grant, David Hubble, Paul Lavrakas, Graham Kalton, Scott Keeter et Courtney Kennedy.

Annexe

Questions sur l'usage du téléphone

National Health Interview Survey

[TRANSCRIPTION]

N1. Existe-t-il dans votre foyer au moins un téléphone qui fonctionne à l'heure actuelle et qui n'est pas un téléphone mobile ?

N2. Est-ce qu'un membre de votre famille possède un téléphone mobile qui fonctionne ?

N3. Combien de téléphones mobiles qui fonctionnent les membres de votre famille possèdent-ils ?

[Si la réponse est « oui » à N1 ainsi que N2, posez N4.]

N4. De tous les appels téléphoniques que reçoit votre famille, est-ce que...

Tous ou presque tous sont reçus sur des téléphones mobiles ?

Certains sont reçus sur des téléphones mobiles et certains sur des téléphones ordinaires ?

Très peu sont reçus sur des téléphones mobiles ou aucun n'est reçu sur des téléphones mobiles ?

California Health Interview Survey – Téléphone mobile

[TRANSCRIPTION]

CC1. Ce téléphone mobile est-il votre seul téléphone ou possédez-vous aussi un téléphone ordinaire à la maison ?

[Si le téléphone est un téléphone mobile et que le répondant possède un téléphone ordinaire, posez la question CC2.]

CC2. De tous les appels téléphoniques que vous recevez, est-ce que...

Tous ou presque tous sont reçus sur des téléphones mobiles ?

Certains sont reçus sur des téléphones mobiles et certains sur des téléphones ordinaires ?

Très peu sont reçus sur des téléphones mobiles ou aucun n'est reçu sur des téléphones mobiles ?

[Si le répondant répond environ la moitié, inscrivez sa réponse.]

California Health Interview Survey – Téléphone fixe

[TRANSCRIPTION]

CL1. Avez-vous un téléphone mobile qui fonctionne ?

[Si la réponse est oui ou que le répondant partage un téléphone mobile, posez CL2.]

CL2. De tous les appels téléphoniques que vous recevez, est-ce que...

Tous ou presque tous sont reçus sur des téléphones mobiles ?

Certains sont reçus sur des téléphones mobiles et certains sur des téléphones ordinaires ?

Très peu sont reçus sur des téléphones mobiles ou aucun n'est reçu sur des téléphones mobiles ?

[Si le répondant répond « environ la moitié », inscrivez sa réponse.]

Pew Research Center for the People & The Press – Téléphone mobile

[TRANSCRIPTION]

PC1. Maintenant, si vous pensez à votre utilisation du téléphone... existe-t-il au moins un téléphone à l'heure actuelle et qui n'est pas un téléphone mobile ?

[Si la réponse est oui, posez PC2.]

ratissage sur des totaux de contrôle démographiques supplémentaires après avoir combiné les deux échantillons.

Étant donné l'état actuel de nos connaissances, nous pensons qu'il existe d'importants avantages à adopter le plan de sondage avec chevauchement complet et y^{ps} avec λ_0 choisi en se basant sur d'autres enquêtes similaires. Il convient de souligner que, même si la CHIS et les enquêtes Pew possédaient des profils de réponse très différents, choisir une valeur de $\lambda_0 = 0,75$ aurait réduit considérablement le biais pour les deux enquêtes. L'un des avantages de cet estimateur par rapport à y^{dep} est, de manière générale, que y^{ps} n'est pas poststratifié en fonction des totaux de domaine d'usage. Nous avons le sentiment que les totaux de domaine d'usage estimés d'après une enquête avec interval sur place (NHIS) peuvent être sujets à des erreurs fort différentes de celles affectant les estimations d'après des enquêtes téléphoniques. Ces différences pourraient donner lieu à des estimations basées sur les enquêtes téléphoniques qui sont biaisées et dont la variance est sous-estimée. Dans le cas des enquêtes au niveau de l'État et au niveau local, pour lesquelles même les totaux pour la situation concernant le type de téléphone utilisé ne sont pas bien connus, les totaux de contrôle pour l'usage du téléphone sont vraisemblablement fort douteux.

Un plan de sondage avec présélection et l'utilisation de y^{scr} comme estimateur à l'avantage de ne nécessiter des totaux de contrôle que pour l'ensemble de la population et pour la composante des utilisateurs d'un téléphone mobile seulement comparables à ceux estimés d'après les données de la NHIS. Un inconvénient est dû au fait que, contrairement aux estimateurs avec chevauchement, il n'existe aucun paramètre de composition pouvant être utilisé pour réduire directement le biais. Le plan avec présélection plus élaboré qui consiste à interviewer les répondants utilisant un téléphone mobile seulement et ceux utilisant un téléphone mobile principalement provenant de la base de numéros de téléphone mobile et à utiliser y^{mod} est valable, mais aucune étude n'a eu pour objectif d'examiner les conditions qui favoriseraient le choix de cet estimateur.

Une analyse plus complexe des effets de l'erreur non due à l'échantillonnage comprendrait d'autres facteurs, tels que l'effet des différences de taux de réponse selon la base de sondage. Par exemple, nous avons noté que les échantillons provenant de la base de numéros de téléphone mobile produisent un plus grand nombre que prévu de ménages utilisant uniquement un téléphone mobile. Ces différences de taux de réponse peuvent être prises en compte dans la répartition de l'échantillon, mais nous ne l'avons pas fait ici. Notre examen de la question révèle que cette prise en compte aboutit à l'affectation d'une plus grande part de l'échantillon à la base de sondage de numéros de téléphone fixe, à l'augmentation de la valeur du facteur de composition et à

un accroissement de l'efficacité des plans avec présélection par rapport aux plans avec chevauchement. Le plan de sondage et l'estimateur avec présélection restent sujets au

biais susmentionné. Bien que la présente étude soit axée sur les erreurs non dues à l'échantillonnage dans les enquêtes téléphoniques à base de sondage double, nous soupçonnons que des problèmes semblables se posent dans le cas de nombreuses autres enquêtes à base de sondage double, mais qu'ils ne sont peut-être pas reconnus. Lohr (2009) mentionne des erreurs non dues à l'échantillonnage dans des enquêtes à base de sondage double générales et suggère de comparer les estimations pour le domaine de chevauchement d'après chaque base de sondage comme simple test diagnostique. Nous pensons qu'il s'agit d'un excellent moyen de commencer l'étude des problèmes associés au chevauchement.

Comme nous l'avons mentionné plus haut, le traitement du chevauchement est un problème important dans les enquêtes à base de sondage double, parce que l'erreur non due à l'échantillonnage peut être liée à la base de sondage. Notre étude montre que les erreurs de non-réponse et de mesure sont reliées à la base de sondage dans les enquêtes téléphoniques à base de sondage double. Il est fort probable que les enquêtes téléphoniques à base de sondage double utilisant des modes différents pourraient être sujettes à des effets analogues. Par exemple, considérons le cas d'une enquête-ménage à base de sondage double conçue pour étudier les membres d'une population rare. Supposons que l'on utilise une liste incomplète des membres de la population avec les numéros de téléphone pour le groupe rare comme base de sondage A et un échantillon probabiliste aréolaire de ménages comme base de sondage B . On pourrait s'attendre que les taux de réponse dans le domaine de chevauchement diffèrent selon la base de sondage et que la différence soit associée aux caractéristiques des répondants, ce qui causera des biais. Même dans le domaine de chevauchement, il pourrait exister des différences telles que celles concernant la durée de l'adhésion de la personne à l'organisation utilisée pour créer la base de sondage A , ce qui pourrait être relié à des caractéristiques comme l'âge. Ce type de situation pourrait présenter un parallèle avec certains problèmes cernés dans le domaine de chevauchement dans le cas des enquêtes téléphoniques. Des erreurs de mesure différentes selon le mode de collecte sont également possibles.

Étant donné la possibilité d'un biais dans les enquêtes à base de sondage double, l'une des constatations importantes de notre étude est que le facteur de composition, λ , influence le biais et a un effet sur la variance. Même si le choix de la valeur de λ n'a habituellement qu'un léger effet sur la variance si cette valeur est proche de la valeur optimale, le biais peut être plus sensible à ce choix. Donc,

quadratique moyenne plus petite que l'approche avec présélection.

Etant donné la possibilité d'un biais dans les enquêtes téléphoniques à base de sondage double pour lesquelles les profils de réponse ressemblent à ceux de la CHIS 2007, nous avons examiné des méthodes d'échantillonnage et d'estimation susceptibles d'être appliquées pour traiter ces biais. Nous avons constaté que les approches avec présélection peuvent être compétitives, et même préférables, pour les enquêtes téléphoniques à base de sondage double quand le biais dû à une non-réponse différentielle ou à l'erreur de mesure est grand. Si le biais n'est pas négligeable, cette constatation est même vraie pour de petites tailles d'échantillon. Toutefois, ces résultats dépendent du choix du facteur de composition et il conviendrait donc de réexaminer la pratique courante consistant à choisir $\lambda = 0,5$. Une autre option consiste à choisir le facteur de composition de manière à éliminer le biais de l'estimateur moyen. Dans de nombreux cas, cette approche peut non seulement éliminer le biais, mais aussi être plus efficace.

Nous avons examiné trois estimateurs en vue de traiter le biais dû à la différence de non-réponse dans le domaine de chevauchement. Le premier est \hat{y}_{ps} , qui utilise la situation concernant le type de téléphone comme variable pour les totaux de contrôle de domaine. Cet estimateur élimine le biais dû à la différence de non-réponse quand λ_0 est utilisé comme facteur de composition, car celui-ci emploie indirectement l'information sur les totaux des domaines d'utilisateurs d'un téléphone fixe principalement et d'un téléphone mobile principalement dans le calcul des taux de réponse selon le domaine et la base de sondage. Un deuxième estimateur, \hat{y}_{sup} , élimine cette source de biais plus directement par poststratification sur les totaux de contrôle pour la situation concernant le type de téléphone et pour l'usage. Cet estimateur permet aussi d'utiliser des facteurs de composition différents dans le domaine de chevauchement afin de réduire la variance des estimations. Le troisième estimateur qui pourrait être utilisé pour réduire le biais est \hat{y}_{mod} , mais cet estimateur est plus pertinent dans le cas d'un plan de sondage qui prévoit l'interview des répondants utilisant un téléphone mobile sélectionné et ceux utilisant un téléphone mobile principalement sélectionnés dans la base de numéros de téléphone mobile, ainsi que tous les répondants provenant de l'échantillon tiré de la base de numéros de téléphone fixe. Ce plan de sondage et cet estimateur avec présélection modifiés pourraient être particulièrement intéressants quant on craint que la moyenne pour les répondants utilisant principalement un téléphone fixe corresponde dans l'échantillon provenant de la base de numéros de téléphone mobile soit sujette à un biais de non-réponse. Tous ces estimateurs pourraient également être calés par

estimer les totaux pour les répondants se servant principalement d'un téléphone mobile. Si nous supposons que $E\hat{y}_{ab}^A(mi) = \bar{Y}_{mi}$ et $E\hat{y}_{ab}^A(mc) = \bar{Y}_{mc}$, il n'est plus nécessaire que $E\hat{y}_{ab}^B(mi) = \bar{Y}_{mi}$ pour que (10) soit sans biais. Comme auparavant, poser que $\lambda_2 = r_1(r_2 - r_1)(r_1 - r_1)^{-1}$ élimine le biais dans l'estimation pour les répondants utilisant principalement un téléphone mobile.

5. Discussion

Le présent examen des erreurs de non-réponse et de mesure dans les enquêtes téléphoniques à base de sondage double donne à penser que les effets de ces erreurs peuvent être très importants. Nous en venons donc à penser que la recherche sur les erreurs non dues à l'échantillonnage en vue de réduire les biais peut être plus importante que celle dominant lieu à des réductions progressives de l'erreur d'échantillonnage.

Les présents travaux de recherche révèlent aussi les lacunes de notre savoir sur les erreurs non dues à l'échantillonnage dans ces enquêtes. La direction et la grandeur des effets de l'erreur de mesure sont particulièrement peu claires. Les incohérences de certains résultats de la CHIS 2007 et des enquêtes Pew pourraient fort bien être dues à des erreurs de mesure associées aux approches différentes de collecte des données suivies dans ces enquêtes ou à des interactions dues aux procédures. Un examen approfondi des sources d'erreur dans les enquêtes téléphoniques à base de sondage double est essentiel si l'on veut améliorer la qualité de ces enquêtes et, à notre avis, des expériences destinées à évaluer les effets des erreurs de mesure seraient particulièrement utiles.

Nous avons constaté que la CHIS 2007 et les enquêtes Pew donnaient lieu à une surreprésentation systématique des utilisateurs d'un téléphone mobile seulement et d'un téléphone mobile principalement dans les échantillons provenant de la base de numéros de téléphone mobile, ainsi qu'une légère surreprésentation des utilisateurs d'un téléphone fixe seulement et d'un téléphone fixe principalement provenant de la base de numéros de téléphone fixe. Cependant, le degré de surreprésentation des domaines varrait selon l'enquête. Dans la CHIS, la surreprésentation aurait pu produire des biais importants dans les estimations si un plan de sondage avec chevauchement et un estimateur moyen simple avaient été adoptés. Une approche de plan de sondage avec présélection a été utilisée pour la CHIS en vue de réduire le biais éventuel, stratégie qui semble avoir été en grande partie couronnée de succès. Dans les enquêtes Pew, la représentation varrait moins selon la base de sondage et le risque de biais était plus faible. Dans ces conditions, l'approche avec chevauchement pourrait produire une erreur

4.2 Approches d'estimation

Une approche proposée par Brick et coll. (2006) consiste à utiliser un plan de sondage avec chevauchement complet ainsi qu'un estimateur moyen pour le chevauchement post-stratifié selon les totaux de domaine d'usage du téléphone, comme dans les enquêtes Pew. Cet estimateur est sans biais et convergent si les estimations dans les domaines sont sans biais et que les tailles d'échantillon de domaine sont suffisamment grandes.

Les données auxiliaires nécessaires pour cette post-stratification pour l'ensemble des États-Unis sont maintenant publiées régulièrement d'après la NHIS. Comme nous l'avons mentionné plus haut, certaines réserves quant à l'utilisation de ces données comme totaux de contrôle méritent d'être examinées plus en détail. Les totaux de contrôle nécessaires pour cet estimateur sont le nombre d'adultes servant d'un téléphone fixe seulement, le nombre d'adultes se servant d'un téléphone mobile seulement, ainsi que les nombres d'adultes qui se servent principalement d'un téléphone fixe et de ceux qui se servent principalement d'un téléphone mobile (N^{ml} et N^{mc} , respectivement). Cela répartit les utilisateurs des deux types de téléphones en deux composantes.

Un autre estimateur du total du domaine de chevauchement en utilisant les mêmes données auxiliaires est donné par

$$\hat{y}^{sep} = \frac{N^a}{N} \hat{y}^a + \frac{N^b}{N} \hat{y}^b + \lambda_1 g_A^{ml} \hat{y}_A^{ab}(ml)$$

$$+ (1 - \lambda_1) g_B^{ml} \hat{y}_B^{ab}(ml) + \lambda_2 g_A^{mc} \hat{y}_A^{ab}(mc) + (1 - \lambda_2) g_B^{mc} \hat{y}_B^{ab}(mc), \quad (7)$$

où les facteurs de poststratification détaillés sont $g_A^{ml} = N^{ml} / \hat{N}^A$, $g_B^{ml} = N^{ml} / \hat{N}^B$, $g_A^{mc} = N^{mc} / \hat{N}^A$, $g_B^{mc} = N^{mc} / \hat{N}^B$, $0 \leq \lambda_1 \leq 1$ et $0 \leq \lambda_2 \leq 1$. Cet estimateur, comme les autres considérés jusqu'à présent, est sans biais et converge en l'absence d'erreurs non dues à l'échantillonnage.

Comme pour l'estimateur (1), les estimations provenant de chaque base de sondage sont poststratifiées avant que l'on calcule leur moyenne. La différence principale entre (1) et (7) est que, dans (7), le domaine des utilisateurs des deux types de téléphone est subdivisé et poststratifié selon l'usage; ce dernier estimateur introduit aussi des facteurs de composition différents dans le domaine de chevauchement.

L'estimateur \hat{y}^{sep} peut être utile quand (1) présente un biais et que les totaux de contrôle de l'usage sont disponibles pour la poststratification. Si les moyennes prévues dans les domaines d'usage sont approximativement égales $(E \hat{y}_B^{ab}(ml) = E \hat{y}_B^{ab}(mc) = E \hat{y}_B^{ab}(mc) = \bar{y}^{ab}(mc) = \bar{y}^{mc})$, alors (7) est sans biais pour tout choix de $0 \leq \lambda_1 \leq 1$ et $0 \leq \lambda_2 \leq 1$. Puisque le biais n'est pas affecté par le choix, différents facteurs de composition peuvent être utilisés pour réduire la variance des estimations comme il est

habituellement proposé de le faire dans la littérature sur les bases de sondage doubles. Le tableau 3 montre que la proportion de répondants dans les domaines d'usage détaillés varie considérablement selon la base de sondage, ce qui pourrait rendre utile l'utilisation de divers facteurs de composition.

Les totaux de contrôle de l'usage du téléphone n'étant souvent pas disponibles, nous avons étudié la modification de (2) afin d'utiliser divers facteurs de composition semblables à ceux employés dans le chevauchement pour (7). Dans ce cas, l'objectif serait de réduire le biais plutôt que la variance. Un estimateur modifié du total dans le domaine de chevauchement est donné par

$$\hat{y}^{mod,ab} = \lambda_1 g_A^{ml} \hat{y}_A^{ab}(ml) + (1 - \lambda_1) g_B^{ml} \hat{y}_B^{ab}(ml) + \lambda_2 g_A^{mc} \hat{y}_A^{ab}(mc) + (1 - \lambda_2) g_B^{mc} \hat{y}_B^{ab}(mc). \quad (8)$$

Cependant, cet estimateur pourrait ne pas être utile pour la réduction du biais. Précédemment, nous avons montré que le biais de $\hat{y}^{ps,ab}$ disparaît quand $\lambda_0 = r(r_c - r_{c1})(r_c^{c1} - r_{c1}^{c1})^{-1}$. Le choix de $\lambda_1 = \lambda_2 = \lambda_0$ élimine le biais à la fois pour les estimations produites pour les personnes se servant principalement d'un téléphone fixe et celles produites pour les personnes se servant principalement d'un téléphone mobile, de sorte que l'emploi de facteurs de composition différents n'aide pas à réduire le biais. Le biais de l'estimateur modifié est donné par

$$b(\hat{y}^{mod,ab}) = W N^{ab} (\bar{X}^{ml} (\lambda_1 r_{l1} r_{l1}^{-1} + (1 - \lambda_1) r_{c1}^{c1} - 1) - \bar{X}^{mc} (\lambda_2 r_{l1} r_{l1}^{-1} + (1 - \lambda_2) r_{c1}^{c1} - 1)), \quad (9)$$

où nous formulons les mêmes hypothèses que celles utilisées plus haut pour approximer le biais de $\hat{y}^{ps,ab}$.

Une autre raison d'étudier un estimateur pour le domaine de chevauchement tel que (8) tient au fait qu'il convient pour les plans de sondage comportant l'élimination par présélection des adultes se servant principalement d'un téléphone fixe provenant de la base de numéros de téléphone mobile. Cette approche a été examinée parce que le nombre de répondants sélectionnés dans la base de numéros de téléphone mobile qui sont classés comme utilisant principalement un téléphone fixe peut être faible et que l'hypothèse selon laquelle $E \hat{y}_B^{ab}(ml) = \bar{y}^{ml}$ risque de ne pas être vérifiée, ce qui pourrait produire un biais.

$$\hat{y}^{mod, \lambda=1,ab} = g_A^{ml} \hat{y}_A^{ab}(ml) + \lambda_2 g_A^{mc} \hat{y}_A^{ab}(mc) + (1 - \lambda_2) g_B^{mc} \hat{y}_B^{ab}(mc). \quad (10)$$

Dans ce plan de sondage, seul l'échantillon de numéros de téléphone fixe est utilisé pour estimer les totaux pour les personnes utilisant un téléphone fixe seulement et celles utilisant principalement un téléphone fixe. Les échantillons provenant des deux bases de sondage sont utilisés pour

Les autres paramètres nécessaires pour la comparaison sont la répartition de la population par domaine selon le type de téléphone utilisé, et nous approximations les valeurs nationales provenant de la NHIS de 2008 ($N_a^* = 0,2N$, $N_b^* = 0,2N$ et $N_{ab}^* = 0,6N$). Dans cette situation, la variance fondée sur un plan de sondage avec chevauchement sous répartition optimale de l'échantillon avec $\lambda = 0,5$ est légèrement plus faible que la variance pour le plan avec présélection sous répartition optimale (le ratio des variances est égal à 0,976). Les variances des deux plans de sondage sont presque les mêmes quand les paramètres de coût sont tels que la présélection des unités provenant de la base de sondage B est un peu moins coûteuse ($c_A^* = 1$, $c_B^* = 3$, $c_s^* = 1,85$).

L'approche de la présélection produit une plus petite erreur quadratique moyenne que l'approche du plan de sondage avec chevauchement sous ces conditions, parce que la première réduit le biais des estimations pour passer de 3-3,3 points de pourcentage à 1,3 point. Même un biais relativement faible domine la comparaison des erreurs quadratiques moyennes entre les deux plans de sondage, en supposant que le biais sous l'approche de présélection est égal à la moitié du biais sous l'approche du chevauchement. Il en est ainsi parce que les variances des plans avec chevauchement et avec présélection sont très similaires. Cependant, si nous utilisons les paramètres provenant des enquêtes Pew, l'erreur quadratique moyenne pour le plan avec chevauchement est plus faible, parce que le biais sous ce plan est plus faible que le biais sous les bases de sondage sous l'approche de présélection.

La répartition entre les bases de sondage sous l'approche avec chevauchement donnée par l'expression (6) en supposant la présence d'une erreur d'échantillonnage seulement est déterminée par les paramètres de population, les paramètres de coût et le facteur de composition. Bien qu'il ne s'agisse pas de la répartition optimale si l'on admet qu'il existe des taux de réponse différents, son examen est néanmoins utile puisqu'il est probable qu'elle soit observée fréquemment en pratique. Dans cette situation, le biais de $y_{ps,ab}$ dû à la non-réponse différentielle peut être éliminé en choisissant λ de manière à satisfaire (4). En se fondant sur les paramètres de la CHIS, la valeur qui élimine ce biais est $\lambda = 0,84$. Si nous poursuivons maintenant les hypothèses de coût et de population susmentionnées, mais en fixant $\lambda = 0,84$, la répartition optimale donnée par (6) donnera alors lieu à la sélection d'environ 75 % de l'échantillon dans la base de numéros de téléphone fixe. Comparativement, si l'on considère la répartition avec $\lambda = 0,5$, 63 % seulement de l'échantillon provient de cette base de sondage. Le choix du facteur de composition est critique. Si l'on utilise $\lambda = 0,84$ conjugué à la répartition optimale pour les paramètres de la CHIS, l'estimateur est sans biais et possède une variance qui est environ 5 % plus faible que celle de l'estimateur sous le plan avec présélection optimal.

$$n_{0,A} = E(C) \tau^{-1} \sqrt{c_A^* N_A^* (N_a^* + \lambda^2 N_{ab}^*)}$$

$$n_{0,B} = E(C) \tau^{-1} \sqrt{c_B^* N_B^* (N_b^* + (1 - \lambda)^2 N_{ab}^*)}, \quad (6)$$

Pour un plan de sondage avec présélection, une fonction de coût linéaire appropriée pour les enquêtes téléphoniques à base de sondage double est donnée par $E(C) = c_A^* n_A + n_B^* c_B^* + N_B^* c_s^* n_b$, où $c_b = c_B + N_B^* c_s^*$, n_b est le nombre échantilloné d'utilisateurs d'un téléphone mobile seulement, et c_s est le coût de la présélection. La variance de l'estimateur avec présélection est $v_{sc}^2 = \sigma^2 (N_A^* n_A^{-1} + N_B^* n_B^{-1})$. La répartition optimale est simplement la répartition stratifiée donnée par $n_{s,A} = E(C) N_A^* (c_A^* N_A^* + \sqrt{c_A^* c_B^*} N_b^*)^{-1}$ et

$$n_B = \frac{E(C) N_B^*}{\sqrt{c_A^* c_B^*} N_A^* + c_B^* N_b^*},$$

$$n_b = \frac{E(C) N_b^*}{\sqrt{c_A^* c_B^*} N_A^* + c_B^* N_b^*}$$

qui produit interviews d'utilisateur d'un téléphone mobile seulement. S'il n'existe pas d'erreur non due à l'échantillonnage et que le coût prévu est fixe, la variance pour le plan avec chevauchement sous répartition optimale est plus faible que la variance pour le plan avec présélection sous répartition optimale quand le coût de la présélection est suffisamment grand pour que $\sqrt{c_b} > N_b^* (\tau - N_A^* \sqrt{c_A^*})$. Si l'on inclut le biais, le plan de sondage avec présélection peut donner une plus petite erreur quadratique moyenne que le plan avec chevauchement, même si cette condition est vérifiée. Dans l'analyse qui suit, nous considérons le biais, mais nous ne tenons pas compte de tous les effets de l'erreur non due à l'échantillonnage. Par exemple, une réponse différentielle influe sur le rendement de la base de sondage dont ont été tirées les unités, ce qui affecte la répartition de l'échantillon et la variance de l'estimation.

Nous comparons les erreurs quadratiques moyennes des plans de sondage avec présélection et avec chevauchement sous les paramètres de la CHIS 2007 donnés précédemment. L'erreur quadratique moyenne est égale à la somme de la variance et du biais au carré. La variance est calculée pour l'estimation globale, mais le biais a seulement pour origine le chevauchement des deux bases de sondage sous nos hypothèses. Les paramètres de coût d'interview et de présélection des unités possédant un téléphone mobile ne sont pas encore bien connus, mais nous utilisons ($c_A^* = 1$, $c_B^* = 3$, $c_s^* = 2$) en nous basant sur l'information donnée par Keeter et coll. (2008) et par Edwards, Brick et Grant (2008).

méthodes appliquées pour la CHIS et les enquêtes Pew peuvent donner lieu à différents échantillons d'adultes.

La source potentielle d'erreur de mesure la plus importante est peut-être liée aux différences entre les questions sur la situation concernant le type de téléphone et l'usage du téléphone utilisées dans les enquêtes. Les questions posées dans chaque enquête sont présentées en annexe. Les différences observées dans les études est due au fait que la CHIS et les enquêtes Pew sont menées par téléphone et que l'on possède des renseignements préalables sur la situation concernant le type de téléphone utilisé.

Les questions employées dans les trois enquêtes sont dérivées de questions utilisées dans un supplément à la Current Population Survey (CPS) de 2004. Comme l'exposent Tucker, Brick et Meekins (2007), des essais cognitifs et de codage du comportement effectués pour le supplément ont suscité certaines réserves concernant les questions de la CPS, surtout la question sur l'usage. Les essais ont révélé que l'absence d'une période de référence précise, l'absence d'un code pour « la moitié du temps » (*half the time*) et la difficulté à fournir la réponse pour d'autres membres du ménage faisaient que la question sur l'usage risquait d'être sujette à une erreur de mesure. Tucker et coll. (2007) soulignent aussi la difficulté qu'avaient les répondants à indiquer la situation concernant le type de téléphone et l'usage pour tous les membres du ménage au moyen d'une seule question. En outre, les répondants avaient de la difficulté à comprendre la signification de « téléphone fixe » (*landline*), « ordinaire » (*regular*), téléphone mobile « qui fonctionne » (*working cell phone*), ainsi que la différence entre utiliser un téléphone mobile et répondre à un téléphone mobile.

Ces questions pourraient avoir une incidence sur la classification par domaine et donc biaiser les estimations. Par exemple, un jeune de 23 ans vivant avec ses parents pourrait déclarer utiliser uniquement un téléphone mobile, tandis que les parents pourraient déclarer utiliser les deux types de téléphone. Les effets de ces types d'erreur de mesure sur les estimations sont difficiles à prédire dans le cas de la NHIS et des enquêtes téléphoniques, mais il n'est pas inattendu d'observer des déclarations non concordantes dans les enquêtes par téléphone et sur place.

Un autre problème de mesure éventuel tient à la relation entre la déclaration concernant l'usage du téléphone et la base de sondage dont proviennent les répondants. L'erreur hypothétique survient si, quand on lui demande quel appareil il utilise pour la plupart de ses appels, le répondant est plus enclin à choisir l'appareil qu'il utilise pour répondre à l'interview. Autant que nous sachions, cette hypothèse n'a pas été testée, mais tout effet d'appareil de cette nature devrait en principe se manifester dans le même sens que l'effet de non-réponse. Un utilisateur des deux types de téléphone devrait

4.1 Approches concernant le plan de sondage

Étant donné les problèmes supplémentaires qui entrent en jeu dans les enquêtes à base de sondage double, les méthodes d'échantillonnage et d'estimation doivent être conçues de manière à tenir compte des sources les plus importantes d'erreur au lieu de se concentrer uniquement sur l'erreur d'échantillonnage. À la présente section, nous examinons les options concernant le plan de sondage et l'estimation pour les enquêtes téléphoniques à base de sondage double dans ces conditions générales de structure d'erreur.

4. Approches d'élaboration du plan de sondage et d'estimation en présence d'erreurs non dues à l'échantillonnage

avoir une plus forte probabilité de déclarer qu'il se sert principalement du téléphone mobile s'il est échantillonné dans la base de numéros de téléphone mobile et être plus susceptible de déclarer qu'il utilise principalement le téléphone fixe s'il est échantillonné dans la base de numéros de téléphone fixe. Donc, le biais dont nous avons discuté plus haut dans le contexte de la non-réponse pourrait être le résultat de l'effet combiné de la non-réponse et de l'effet d'appareil. S'il n'est pas possible de déterminer l'importance de ces sources de biais, le choix de méthodes pour réduire ce biais n'est pas clair.

Une décision essentielle concernant le plan de sondage double est celle de savoir si l'on doit utiliser un plan de sondage avec présélection ou avec chevauchement complet. Nous commençons par examiner la répartition optimale de l'échantillon pour les plans avec chevauchement et avec présélection appropriés pour les enquêtes téléphoniques à base de sondage double quand des échantillons aléatoires simples sont tirés indépendamment des deux bases de sondage et que $N_a > 0$, $N_b > 0$, et $N_{ab} > 0$. Nous supposons tout au long de l'exposé que les tailles d'échantillon sont suffisamment grandes pour que l'on puisse ignorer les facteurs de correction pour population finie.

Nous utilisons une fonction de coût prévu linéaire $E(C) = c_A(n_A + n_B c_B c_A^-)$, où c_A est le coût d'une interview par téléphone fixe, c_B est le coût d'une interview par téléphone mobile, et n_A et n_B sont les nombres d'unités échantillonnées dans les bases de sondage A et B , respectivement. Si nous supposons que la variance élémentaire, σ^2 , est constante, la variance de l'estimateur avec chevauchement est $v_{ov}^2 = \sigma^2(N_A(N_A + \lambda^2 N_{ab})n_A^{-1} + N_B(N_B + (1 - \lambda)^2 N_{ab})n_B^{-1})$. La répartition de l'échantillon qui minimise la variance en utilisant cette fonction de coût peut être trouvée en se servant de méthodes Lagrangiennes classiques et est

variabilité des taux de réponse est plus grande dans la base de numéros de téléphone mobile.

En raison du biais éventuel dans le plan de sondage avec chevauchement, Brick et coll. (2006) proposent d'utiliser un plan de sondage avec présélection conçu pour exclure les adultes appartenant à des ménages utilisant les deux types de téléphone s'ils ont été échantillonnés dans la base de numéros de téléphone mobile. Dans un plan avec présélection, un biais persiste à cause de la différence de non-réponse chez les utilisateurs des deux types de téléphone selon l'usage du téléphone dans l'échantillon de numéros de téléphone fixe. Si nous substituons $\lambda = 1$ dans (2) et (3), le biais de $y_{scr,ab}^A = g^A y_{ab}^A$ est donné par

$$b(\hat{y}_{scr,ab}^A) = W N^{ab}(\bar{Y}_{mi}^A - \bar{Y}_{mi}^A)(t_1^A t_1^A - 1). \tag{5}$$

Le biais pour ce plan de sondage et cet estimateur équivalait à celui des estimateurs pour base de sondage unique, et il disparaît si $\bar{Y}_{mi}^A = \bar{Y}_{mi}^A$ ou que les taux de réponse dans l'échantillon de numéros de téléphone fixe sont les mêmes pour les utilisateurs se servant principalement du téléphone fixe et ceux se servant principalement du téléphone mobile. Soulignons que, sous ce plan de sondage, aucun facteur de composition ne peut être utilisé pour contrôler le biais.

Le biais de l'estimateur avec présélection pour la CHIS 2007 est à peu près égal à la moitié de celui de l'estimateur moyen en utilisant $\lambda = 0,50$ (le biais de présélection est de 1,3 point de pourcentage comparativement à celui de l'estimateur moyen poststratifié en utilisant $\lambda = 0,50$ qui est de -3,3 points). Pour les paramètres des enquêtes Pew, le biais de l'estimateur moyen poststratifié et celui de l'estimateur avec présélection sont presque égaux, le biais de l'estimateur avec présélection étant légèrement plus grand que celui de l'estimateur poststratifié (le biais pour l'estimateur avec présélection est de 1,1 point de pourcentage comparativement à -0,7 point pour le chevauchement poststratifié).

Un problème mentionné plus haut tient au fait que les totaux de domaine pour la poststratification, même si l'on ne considère que la situation concernant le type de téléphone (domaines des utilisateurs du téléphone fixe seulement, du téléphone mobile seulement et des deux types de téléphone) ne sont généralement pas disponibles pour les enquêtes au niveau de l'État ou au niveau local. Des estimations sur petits domaines du pourcentage d'adultes qui utilisent un téléphone mobile seulement ont été publiées au niveau de l'État (Blumberg, Luke, Davidson, Davern, Yu et Soderberg 2009), mais ces auteurs ne fournissent pas d'estimations sur les totaux de contrôle de l'usage du téléphone, la situation est encore plus limitée, seules les estimations nationales de la NHIS étant publiées. Puisque, pour la base de sondage

des numéros de téléphone mobile, les taux de réponse varient habituellement selon l'usage, certaines hypothèses au sujet des taux de réponse dans l'échantillon de numéros de téléphone mobile peuvent être utiles pour éviter une sureprésentation importante des adultes utilisant seulement ou utilisant principalement un téléphone mobile dans l'échantillon tiré de la base de sondage de numéros de téléphone mobile quand on utilise le plan de sondage avec chevauchement.

3.2 Effets des erreurs de mesure

En plus de la non-réponse, certaines différences entre les distributions observées dans les tableaux 1 et 2 pourraient être dues à une erreur de mesure. Avant de parler des hypothèses associées à l'erreur de mesure, nous discutons de certaines procédures importantes des enquêtes qui pourraient être reliées à cette erreur. Il s'agit de différences fondamentales entre les enquêtes, telles que le mode de collecte et le thème. La NHIS est une enquête avec interview sur place, tandis que la CHIS et les enquêtes Pew sont des enquêtes téléphoniques. La NHIS ainsi que la CHIS couvrent une grande gamme de sujets.

Ces enquêtes s'appuient aussi sur des méthodes différentes pour recueillir l'information sur la situation en ce qui concerne le type de téléphone ainsi que sur l'usage du téléphone. Dans la NHIS, on demande à un membre adulte de la famille de répondre à des questions sur la situation concernant le type de téléphone et sur l'usage du téléphone pour la famille complète durant une partie de l'interview réservée aux caractéristiques de la famille. Dans le cas de l'échantillon de numéros de téléphone mobile de la CHIS 2007, des questions sur la situation concernant le type de téléphone sont posées durant l'interview de présélection du ménage, mais les questions sur l'usage du téléphone sont posées durant l'interview de l'adulte sélectionné pour participer à l'enquête. Dans le cas de l'échantillon de numéros de téléphone fixe de la CHIS et des enquêtes Pew, les questions sur la situation concernant le type de téléphone et sur l'usage du téléphone sont toutes regroupées dans l'une des dernières parties de l'interview de la personne adulte sélectionnée. Ce placement des questions à un moment plus tardif est possible parce qu'aucune présélection n'est effectuée. L'échantillonnage d'un adulte est une autre procédure qui peut interagir avec le processus de mesure. Dans la CHIS 2007, un adulte est sélectionné parmi tous ceux partageant le même numéro de téléphone mobile. Dans les enquêtes Pew, et dans la plupart des autres enquêtes par téléphone mobile, ce dernier est considéré comme un appareil personnel et la personne qui y répond est interviewée. Dans les ménages utilisant les deux types de téléphone, les

Tableau 3
Facteurs de poststratification relatifs dans le domaine de
chevauchement pour la CHIS 2007 et les enquêtes Pew

Facteurs de poststratification relatifs*	CHIS 2007	Enquêtes Pew
$r_{11}^{-1} \div g_A / g_{ml}$	1,09	1,07
$r_{12}^{-1} \div g_A / g_{mc}$	0,50	0,84
$r_{21}^{-1} \div g_B / g_{ml}$	0,74	0,78
$r_{22}^{-1} \div g_B / g_{mc}$	2,42	1,51

Facteur d'ajustement par poststratification pour le domaine d'usage du téléphone dans le domaine de chevauchement divisé par le facteur de poststratification du domaine de chevauchement.

$\bar{Y}_i - \bar{Y}_{mc}$ sont nécessaires pour des caractéristiques étudiées au moyen d'une enquête téléphonique à base de sondage double plutôt qu'en formulant des hypothèses arbitraires comme dans l'exemple qui précède. Blumberg et Luke (2009) donnent des estimations qui donnent à penser que ces différences peuvent être aussi importantes que les celles entre la population d'utilisateurs de téléphone mobile seulement et d'utilisateurs de téléphone fixe qui ont été décrites en détail ailleurs. Cependant, les estimations de la NHIS sont calculées au moyen de données venant d'une enquête avec interview sur place et non d'une enquête téléphonique à base de sondage double.

Keeter, Dimock et Christian (2008) donnent des estimations des caractéristiques des utilisateurs des deux types de téléphone selon la base de sondage, mais pas de façon suffisamment détaillée pour pouvoir calculer les biais. Les données de Keeter indiquent que les estimations pour les utilisateurs doubles sélectionnés dans la base de numéros de téléphone mobile pourraient être plus proches des estimations de la NHIS pour le domaine de chevauchement que les estimations pour les utilisateurs doubles tirés de la base de numéros de téléphone fixe. Cependant, puisque les taux de réponse dans le domaine de chevauchement sont plus variables pour les unités provenant de la base de numéros de téléphone mobile que pour celles provenant de la base de téléphone fixe, un plan de sondage avec présélection visant à réduire le biais exclure les utilisateurs doubles provenant de la base de numéros de téléphone mobile plutôt que de la base de numéros de téléphone fixe quand la

La condition (a) est essentiellement la condition bien connue issue de la méthode à base unique. La condition (b) diffère des expressions pour une base de sondage unique parce que le biais dépend à la fois des taux de réponses relatifs et du facteur de composition, λ . Une exception a lieu quand $r_1^1 r_1^{-1} = r_1^2 r_1^{-2}$, ou de manière équivalente $r_1^1 r_1^{-1} = r_1^2 r_1^{-2}$, où r_1^2 est le taux de réponse des unités utilisant principalement le téléphone mobile dans l'échantillon de numéros de téléphone fixe et r_1^2 est le taux de réponse des unités utilisant principalement le téléphone mobile dans l'échantillon de numéros de numéros de téléphone mobile. Sous cette forme, l'expression est comparable à celle du biais pour une base de sondage unique qui montre que le biais est nul quand les taux de réponse sont constants. De façon plus générale, la valeur de λ influe non seulement sur la variance de l'estimation, mais aussi sur son biais. Celui-ci peut être éliminé en choisissant

Puisque la proportion du total de population couverte par la base de numéros de téléphone fixe est approximativement égale à la proportion couverte par la base de numéros de téléphone mobile, la valeur de $\lambda = 0,50$ a été utilisée dans la plupart des applications sans tenir compte de l'effet sur le

$$(4) \quad \gamma^0 = \frac{1/\mu - 1/\mu^2}{(1/\mu^2 - 1/\mu^3)}.$$

...bais.

Nous pouvons maintenant appliquer ces expressions pour évaluer le biais de l'estimateur sous enquête téléphonique à base de sondage double dans le cas de la CHIS, en supposant que ce biais est dû seulement à la différence de non-réponse dans le domaine de chevauchement. Si nous utilisons les données du tableau 1, $W = 0,74$ pour la région Ouest selon la NHIS. Nous approximations $r_1 r_1^{-1}$ par le facteur de poststratification relatif qui est le ratio du pourcentage d'unités de l'échantillon de numéros de téléphone fixe de la CHIS classées comme utilisant principalement le téléphone fixe au pourcentage d'adultes participant à la NHIS dans les ménages ayant un téléphone fixe qui utilisent principalement le téléphone fixe; nous calculons de la même façon $r_{c1} r_{c1}^{-1}$ pour les quantités ayant trait au téléphone mobile. Les quantités estimées pour la CHIS 2007 sont présentées au tableau 3, $r_1 r_1^{-1} = 1,09$ pour l'échantillon de numéros de téléphone fixe et $r_{c1} r_{c1}^{-1} = 0,50$ pour l'échantillon de numéros de téléphone mobile. À titre d'exemple, si nous supposons que $\bar{Y}_{nl} = 0,3$ et $\bar{Y}_{mc} = 0,5$, le biais du pourcentage estimé basé sur (3) est de l'ordre de 3 points de pourcentage (un biais relatif d'environ 9 %) si $\lambda = 0,5$. En utilisant (4), le biais est nul quand $\lambda = 0,84$; le biais devient négatif pour les valeurs plus grandes de λ . Les mêmes calculs peuvent être effectués en utilisant les données provenant des enquêtes Pew, et les estimations sont également présentées au tableau 3. Les paramètres diffèrent

Les répartitions des réponses selon la base de sondage et selon l'usage du téléphone que l'on observe pour ces enquêtes corroborent, dans les deux cas, la conjecture concernant l'accessibilité de Brick et coll. (2006). Cette conjecture implique un classement des répondants selon le degré d'accessibilité et la probabilité de répondre – dans la base de numéros de téléphone mobile, le classement allant des personnes les plus susceptibles de répondre à celles les moins susceptibles de le faire est le suivant : téléphone mobile seulement, principalement le téléphone mobile et principalement le téléphone fixe. Le problème particulier que pose l'utilisation de deux bases de sondage est que le classement dans la base des numéros de téléphone fixe est différent (téléphone fixe seulement, principalement téléphone fixe et principalement téléphone mobile) et que les unités comprises dans le domaine de chevauchement des deux bases de sondage pourraient avoir des taux de réponse et des biais très différents.

Pour examiner le biais de non-réponse dans le cas d'une enquête à base de sondage double avec chevauchement, supposons que l'on poststratifie les échantillons de numéros de téléphone fixe et de numéros de téléphone mobile selon les totaux de domaine pour la situation concernant le type de téléphone avant de produire une estimation globale moyenne. L'estimateur poststratifié est donné par

$$\hat{y}^{ps} = \frac{N}{N^a} \hat{y}^a + \frac{N}{N^b} \hat{y}^b + \lambda g^a \hat{y}^a + (1 - \lambda) g^b \hat{y}^b, \quad (1)$$

où le facteur de poststratification est N^a / \hat{N}^a pour l'échantillon de numéros de téléphone fixe seulement, N^b / \hat{N}^b pour l'échantillon de numéros de téléphone mobile seulement, et les facteurs de poststratification propres à la base de sondage pour le domaine de chevauchement sont $g^a = N^{ab} / \hat{N}^a$ et $g^b = N^{ab} / \hat{N}^b$ pour les échantillons de numéros de téléphone fixe et de téléphone mobile, respectivement. Les estimateurs d'Horvitz-Thompson (HT) du nombre d'unités sont \hat{N}^a pour le domaine des téléphones fixes seulement, \hat{N}^b pour le domaine des téléphones mobiles seulement, et \hat{N}^{ab} et \hat{N}^B pour le domaine de chevauchement des deux échantillons. Puisque nous nous concentrons sur le chevauchement, nous écrivons

$$\hat{y}^{ps,ab} = \lambda g^a \hat{y}^a + (1 - \lambda) g^b \hat{y}^b. \quad (2)$$

Si nous tenons compte, dans le domaine de chevauchement, de différences de taux de réponse selon l'usage du

téléphone telles que celles observées dans les enquêtes téléphoniques à base de sondage double, l'estimateur (2) est biaisé. Soit W la proportion du domaine de chevauchement correspondant à des unités utilisant le téléphone fixe principalement, et soit \bar{Y}^{ml} et \bar{Y}^{mc} les moyennes de population d'une caractéristique pour les utilisateurs doubles se servant principalement du téléphone fixe (désignés par ml pour l'anglais *mainly land*) et ceux se servant principalement du téléphone mobile (désignés par mc pour l'anglais *mainly cell*), respectivement. Le biais de $\hat{y}^{ps,ab}$ est

$$b(\hat{y}^{ps,ab}) = W N^{ab} (\bar{Y}^{ml} - \bar{Y}^{mc})$$

$$(\lambda r_1 r_1^{-1} + (1 - \lambda) r_1^{cl} r_1^{-1} - 1), \quad (3)$$

où r_1 est le taux de réponse des utilisateurs doubles pour l'échantillon de numéros de téléphone fixe, r_1^{cl} est le taux de réponse des utilisateurs doubles se servant principalement du téléphone fixe dans l'échantillon de numéros de téléphone fixe, r_2 est le taux de réponse des utilisateurs doubles dans l'échantillon de numéros de téléphone mobile, et r_1^{cl} est le taux de réponse des utilisateurs doubles se servant principalement du téléphone fixe dans l'échantillon de numéros de téléphone mobile.

Pour dériver (3), nous commençons par définir les estimateurs de domaine comprenant les utilisateurs doubles se servant principalement du téléphone fixe et se servant principalement du téléphone mobile provenant de l'échantillon de numéros de téléphone fixe comme étant $\hat{y}^{ab}(ml) = \hat{N}^{ml} \hat{y}^{ab}(ml)$ et $\hat{y}^{ab}(mc) = \hat{N}^{mc} \hat{y}^{ab}(mc)$, ceux et provenant de l'échantillon de numéros de téléphone mobile comme étant $\hat{y}^B(ml) = \hat{N}^B \hat{y}^B(ml)$ et $\hat{y}^B(mc) = \hat{N}^B \hat{y}^B(mc)$. Maintenant, supposons que a) $E \hat{y}^{ab}(ml) = E \hat{y}^{ab}(mc) = \bar{Y}^{ml}$ et $E \hat{y}^{ab}(mc) = E \hat{y}^{ab}(ml) = \bar{Y}^{mc}$; b) les covariances sont telles que $\text{cov}(\hat{N}^{ml} / \hat{N}^{ab}, \hat{y}^{ab}(ml)) = 0$; et c) les totaux de domaine prévus sont des expressions simples telles que $E \hat{N}^{ml} = r_1 N^{ml}$, $E \hat{N}^{ab} = r_1 N^{ab}$, etc. Puisque $E(\hat{N}^{ab} / \hat{N}^{ab}) = \hat{y}^{ab} = N^{ab} E \{ (\hat{N}^{ml} \hat{y}^{ab}(ml) + \hat{N}^{mc} \hat{y}^{ab}(mc)) / \hat{N}^{ab} \}$, nous pouvons écrire $E(\hat{N}^{ab} / \hat{N}^{ab}) \hat{y}^{ab} = r_1 r_1^{-1} N^{ml} \bar{Y}^{ml} + r_1^{cl} r_1^{-1} N^{mc} \bar{Y}^{mc} = N^{ab} (r_1 r_1^{-1} W (\bar{Y}^{ml} - \bar{Y}^{mc}) + \bar{Y}^{mc})$. Une expression correspondante peut être écrite pour $E g^B \hat{y}^B$. La combinaison des deux donne (3).

Ces expressions supposent que $E \hat{y}^{ab}(ml) = \bar{Y}^{ml}$ et $E \hat{y}^B(mc) = \bar{Y}^{mc}$. Une autre approche qui ne nécessite pas cette hypothèse consiste à poser qu'il existe une association entre la propension à répondre et l'usage du téléphone. Dans ce cas, le biais serait une fonction de la propension à répondre des unités provenant de chaque base de sondage. Nous n'examinons pas l'approche de la propension à répondre ici.

L'expression (3) montre que, quand $0 < W < 1$, le biais de $\hat{y}^{ps,ab}$ est nul si a) $\bar{Y}^{ml} = \bar{Y}^{mc}$; ou b) $\lambda r_1 r_1^{-1} + (1 - \lambda)$

c'est-à-dire une période correspondant approximativement à la période de collecte des données de la CHIS. Les chiffres de la CHIS correspondent aux proportions non pondérées dans l'échantillon (les proportions pondérées sont presque identiques). Bien qu'une approche de présélection ait été utilisée pour la CHIS, l'information sur l'usage du téléphone a été recueillie pour chaque ménage répondant compris dans l'échantillon de téléphones mobiles. Le tableau montre que, comparativement aux estimations de la NHIS, la répartition dans la base de numéros de téléphone mobile surreprésente le pourcentage d'adultes faisant partie de ménages dotés uniquement d'un téléphone mobile et sous-représente ceux faisant partie de ménages utilisant principalement le téléphone fixe. Les répondants de l'échantillon de numéros de téléphone fixe surreprésentent les utilisateurs d'un téléphone fixe uniquement et sous-représentent les utilisateurs doubles se servant principalement d'un téléphone mobile. Les différences pour la base de numéros de téléphone fixe sont plus importantes que celles observées dans une enquête menée

en 2004 dont les résultats sont présentés dans Brick et coll. (2006). Le tableau 2 donne le même type de comparaison des estimations nationales d'après la NHIS pour la deuxième moitié de 2008 aux résultats non pondérés des enquêtes Pew agrégées (toutes les enquêtes ont été réalisées auprès d'un échantillon à probabilités d'inclusion égales). Comme dans le cas de la CHIS, la répartition dans la base de numéros de téléphone mobile utilisée pour les enquêtes Pew surreprésente le pourcentage dans le groupe possédant un téléphone mobile seulement et sous-représente celui dans le groupe d'utilisateurs doubles se servant principalement d'un téléphone fixe, mais les différences sont moins marquées que dans le cas de la CHIS. La répartition d'après l'échantillon de numéros de téléphone fixe des enquêtes Pew correspond de près à celle observée pour la NHIS, avec une légère sous-représentation du groupe d'utilisateurs doubles se servant principalement d'un téléphone mobile.

Tableau 1
Répartition en pourcentage des adultes selon l'usage du téléphone d'après la CHIS de 2007 et la NHIS

Usage du téléphone		NHIS Ouest – Adultes dans les ménages à téléphone fixe		CHIS 2007 – Base de numéros de téléphone fixe		NHIS Ouest – Adultes dans les ménages à téléphone mobile		CHIS 2007 – Base de numéros de téléphone mobile	
Téléphone fixe seulement		23,5 %	34,2 %			—		—	
Double – principalement fixe		(1,5 %) 56,6 %	(0,2 %) 53,2 %			60,9 %		18,5 %	
Double – principalement mobile		(1,7 %) 19,9 %	(0,2 %) 12,7 %			(1,7 %) 21,4 %		(0,7 %) 31,2 %	
Téléphone mobile seulement		—	—			(1,4 %) 17,7 %		(0,9 %) 50,3 %	
Total		100,0 %	100,0 %			100,0 %		100,0 %	

Notes : NHIS-Ouest désigne la National Health Interview Survey, région de l'Ouest, six premiers mois de 2008, avec les pourcentages de tous les ménages possédant le type de service particulier (merci à S. Blumberg et J. Luke pour cette totalisation spéciale). CHIS 2007 représente la California Health Interview Survey, dont les données ont été recueillies en 2007 et au début de 2008, avec les pourcentages non pondérés d'après les bases de numéros de téléphone fixe et de numéros de téléphone mobile. Dans l'échantillon de téléphones mobiles, l'information sur l'usage a été obtenue durant l'interview de présélection. Les erreurs-types approximatives figurent entre parenthèses.

Tableau 2
Répartition en pourcentage des adultes selon l'usage du téléphone d'après les enquêtes Pew et la NHIS

Usage du téléphone		NHIS – Adultes dans les ménages à téléphone fixe		Enquêtes Pew – Base de numéros de téléphone fixe		NHIS – Adultes dans les ménages à téléphone mobile		Enquêtes Pew – Base de numéros de téléphone mobile	
Téléphone fixe seulement		19,4 %	23,0 %			—		—	
Double – principalement fixe		(0,7 %) 58,8 %	(0,4 %) 62,7 %			58,8 %		42,3 %	
Double – principalement mobile		(0,8 %) 19,3 %	(0,5 %) 14,4 %			(0,8 %) 18,5 %		(0,8 %) 24,0 %	
Téléphone mobile seulement		—	—			(0,7 %) 22,7 %		(0,7 %) 33,7 %	
Total		100,0 %	100,0 %			100,0 %		100,0 %	

Notes : NHIS désigne la National Health Interview Survey, six derniers mois de 2008, avec les pourcentages de tous les ménages possédant le type de service particulier. Les enquêtes Pew regroupent huit enquêtes réalisées pour le Pew Research Center for the People & the Press d'octobre 2008 à mars 2009, avec les pourcentages non pondérés provenant des bases de numéros de téléphone fixe et de numéros de téléphone mobile. (Merci à S. Keeter d'avoir fourni ces données). Les erreurs-types approximatives figurent entre parenthèses.

sondage, surtout si les méthodes de collecte de données diffèrent selon la base de sondage. Troisièmement, l'échantillonnage à partir de plus d'une base de sondage accroît la complexité et crée un plus grand nombre de situations dans lesquelles les erreurs non dues à l'échantillonnage pourraient avoir des effets différentiels.

3.1 Effets de la non-réponse

Brick, Dippo, Presser, Tucker et Yuan (2006) montrent que la surreprésentation du nombre d'adultes appartenant à un ménage muni uniquement d'un téléphone mobile qui se produit dans presque tous les échantillons d'enquête téléphonique à base de sondage double peut être due à l'erreur de non-réponse. Selon ces auteurs, la surreprésentation pourrait être le résultat d'une différence d'accessibilité, en ce sens que les adultes qui utilisent rarement leur téléphone mobile sont moins susceptibles de répondre à un appel sur celui-ci que ceux qui l'utilisent régulièrement. Ils n'ont pas constaté le même genre de différence de taux de réponse liée à l'usage dans l'échantillon de ménages dotés d'un téléphone fixe. Kennedy (2007) explore plus en détail ce type de biais de non-réponse en examinant les effets sur des estimations particulières.

Afin d'évaluer la différence de représentation, nous comparons les répartitions des échantillons de la CHS 2007 et des enquêtes Pew selon la base de sondage et l'usage du téléphone à des estimations provenant de la National Health Interview Survey (NHIS). La NHIS est une enquête par interview sur place parrainée par le National Center for Health Statistics dont la collecte des données est effectuée par le U.S. Bureau of Census (les données de la NHIS ont été fournies par S. Blumberg et J. Luke sous forme de totales spéciales). Il s'agit de la seule enquête du gouvernement fédéral qui fournit des estimations sur la situation concernant le type de téléphone et l'usage de celui-ci (Blumberg et Luke 2009). Nous définissons l'usage pour les utilisateurs doubles (les membres des ménages dotés des deux types de service téléphonique) comme étant principalement mobile ou principalement fixe, la catégorie principalement mobile comprenant les personnes qui vivent dans les ménages qui reçoivent tous ou presque tous leurs appels sur leur téléphone mobile, et la catégorie principalement fixe comprenant les utilisateurs doubles dans les ménages qui ne reçoivent pas tous ou presque tous leurs appels sur leur téléphone mobile.

Pour que les chiffres soient plus comparables à ceux de la CHS, dans le tableau 1, nous limitons les estimations de la NHIS à celles obtenues pour la région de l'Ouest seulement (les estimations de la NHIS pour la Californie ne sont pas disponibles). La Californie comprend 52 % des adultes vivant dans l'Ouest. Les chiffres de la NHIS sont des chiffres de population pour les six premiers mois de 2008,

enquêtes sont réalisées auprès de l'ensemble de la population adulte des États-Unis. Elles comprennent l'interview d'un adulte dans chaque ménage échantillonné dans l'une ou l'autre base de sondage en utilisant des questionnaires presque identiques. Pour l'ensemble de huit enquêtes, près de 11 300 interviews par téléphone fixe et 3 800 interviews par téléphone mobile ont été effectuées. Les taux de réponse aux diverses enquêtes sont fort semblables pour les échantillons de numéros de téléphone fixe et de téléphone mobile, avec un écart médian d'un point de pourcentage entre les échantillons provenant des deux bases de sondage. Sur l'ensemble de huit enquêtes et des deux bases de sondage, les taux de réponse varient entre 17 % et 24 %.

Dans les enquêtes Pew, comme dans la plupart des enquêtes téléphoniques à base de sondage double avec chevauchement, une version calée de l'estimateur moyen est employée pour produire les estimations. Pour la plupart des enquêtes, le calage est effectué à la fois sur les dénombrements de domaine pour le type de téléphone (nombres respectifs d'adultes vivant dans un ménage ne possédant qu'un téléphone mobile, dans un ménage ne possédant qu'un téléphone fixe et dans un ménage possédant à la fois un téléphone fixe et un téléphone mobile) ainsi que sur des variables démographiques. Pour les enquêtes Pew, le calage est également effectué sur des totaux de variables démographiques, dont l'âge, le niveau de scolarité, la race/l'ethnicité, la région et la densité de population pour les ménages comptant des adultes de 18 ans et plus. En outre, un calage est effectué sur les totaux selon le type de téléphone et, dans le domaine de chevauchement, selon l'usage relatif des téléphones fixe et mobile.

3. Erreurs non dues à l'échantillonnage

La théorie des enquêtes à base de sondage double a été élaborée sous des conditions idéales de réponse complète et d'absence d'autres erreurs non dues à l'échantillonnage. Les erreurs non dues à l'échantillonnage ont une incidence sur le biais et sur la précision des estimations de toute enquête, mais, pour trois raisons, leurs effets dans les enquêtes à base de sondage double peuvent différer qualitativement de ceux dans les enquêtes à base de sondage simple. Premièrement, dans les enquêtes à base de sondage double, l'erreur non due à l'échantillonnage rend souvent difficile la détermination de la probabilité de sélection de l'unité échantillonnée. Cela se produit quand l'appartenance au domaine est confirmée durant la collecte de données et que la non-réponse et les erreurs de mesure font qu'il est difficile de déterminer si l'unité échantillonnée se trouve dans le domaine de chevauchement. Deuxièmement, dans les enquêtes à base de sondage double, l'erreur non due à l'échantillonnage peut être reliée directement, parfois de manière causale, à la base de

Les données provenant de la CHIS 2007 sont utilisées pour illustrer les problèmes que pose une enquête téléphonique à base de sondage double s'appuyant sur une approche de présélection. La CHIS 2007 est une enquête téléphonique ayant pour cible la population californienne qui a été réalisée par le UCLA Center for Health Policy Research en collaboration avec le California Department of Public Health, le California Department of Health Care Services et le Public Health Institute. La collecte des

2.3 Applications aux enquêtes téléphoniques

D'autres méthodes d'estimation qui ont été prises en considération pour une enquête avec chevauchement comprennent l'estimateur pour base de sondage simple (Bankier 1986; Kalton et Anderson 1986; Skinner 1991) et l'estimateur du pseudo-maximum de vraisemblance (Skinner et Rao 1996; Lohr et Rao 2000; Lohr et Rao 2006). Lohr (2009) passe ces estimateurs en revue. Presque toutes les enquêtes téléphoniques avec chevauchement que nous avons vues utilisent certaines versions de l'estimateur moyen, qui est le point de concentration de la présente étude.

L'estimation de la variance en se servant de l'estimateur moyen est relativement simple si λ est fixe et indépendant de l'échantillon sélectionné. Dans ce cas, $V(\hat{y}^{ave}) = V(\hat{y}^a + \lambda \hat{y}^b_{ab}) + V(\hat{y}^b + (1 - \lambda) \hat{y}^b_{ab})$, et chacune de ces variances peut être calculée en utilisant des méthodes d'estimation de la variance appropriées pour les échantillons distincts. Si λ dépend de l'échantillon, comme dans le cas des estimateurs de Hartley et de Fuller et Burneister, l'estimation de la variance est plus compliquée. Les estimateurs moyens avec un facteur λ fixe ont été utilisés dans la plupart des enquêtes téléphoniques à base de sondage double avec chevauchement. Nous discutons de cette approche plus loin pour les enquêtes Few.

est donné par $\hat{y}^{ave} = \hat{y}^a + \hat{y}^b + \lambda \hat{y}^a_{ab} + (1 - \lambda) \hat{y}^b_{ab}$, avec $0 \leq \lambda \leq 1$. À l'instar de Lohr (2009), nous donnons à ces estimateurs le nom d'estimateurs moyens. Si nous supposons que \hat{y}^a et \hat{y}^b sont sans biais pour le domaine a et pour le domaine b , respectivement et que \hat{y}^a_{ab} et \hat{y}^b_{ab} sont tous deux sans biais pour le domaine ab , alors \hat{y}^{ave} est un estimateur sans biais du total. Les estimations des moyennes et d'autres quantités peuvent être produites en utilisant des pondérations, celles pour les unités comprises dans ab tirées de A étant multipliées par λ et celles pour les unités tirées de B étant multipliées par $(1 - \lambda)$. Le choix du facteur de composition, λ , a été étudié par de nombreux chercheurs et des valeurs particulières en vue de réduire la variance des estimations ont été proposées par Hartley (1962, 1974) et par Fuller et Burneister (1972). Tous les estimateurs moyens requièrent que le domaine puisse être précisé pour chacune des unités échantillonnées.

Dans le cas des enquêtes téléphoniques à base de sondage double avec chevauchement, nous utilisons des données agrégées provenant de huit enquêtes réalisées par le Pew Research Center for the People & the Press de la fin de 2008 au début de 2009. (Les données provenant des enquêtes Pew ont été fournies par Scott Keeter, du Pew Research Center for the People & the Press). Toutes ces

Dans la CHIS 2007, les estimations pour l'échantillon de numéros de téléphone mobile sont calées sur la population californienne adulte possédant un téléphone mobile à l'étape de la présélection (avant la correction de la pondération pour tenir compte de la non-réponse de l'adulte échantillonné). L'obtention de totaux de contrôle fiables pour le calage au niveau de l'État pose certaines difficultés dont il sera question plus loin. Les échantillons tirés des deux bases de sondage sont indépendants et traités en tant que tels jusqu'à la dernière étape, où ils sont combinés et calés sur des totaux indépendants pour l'ensemble de la population californienne adulte. Cette dernière étape de calage n'inclut pas le type de téléphone comme domaine.

Dans l'échantillon de numéros de téléphone fixe de la CHIS 2007, on a échantillonné et interviewé un adulte par ménage. Dans l'échantillon de numéros de téléphone mobile, les personnes vivant dans un ménage possédant un numéro de téléphone fixe ont été écartées; un adulte par ménage a été échantillonné et interviewé dans l'échantillon de numéros de téléphone mobile s'il vivait dans un ménage classé comme possédant uniquement un téléphone mobile. Tous les ménages répondants, y compris ceux éliminés de la base de numéros de téléphone mobile, ont répondu à des questions sur la situation concernant les types de téléphone et sur leur usage. Près de 49 000 adultes provenant de l'échantillon de numéros de téléphone fixe ont été interviewés, ainsi que 825 adultes possédant uniquement un téléphone mobile. Pour l'échantillon de numéros de téléphone fixe, le taux de réponse était de 35,5 % à l'interview menée auprès d'une personne fournissant des renseignements sur le ménage et de 59,4 % à l'interview auprès de l'adulte échantillonné. Pour la base de numéros de téléphone mobile, les taux de réponse correspondants étaient de 22,1 % et de 52,0 %. Puisque la CHIS 2007 a été réalisée en adoptant une approche de présélection, le taux de réponse publié pour l'interview des personnes fournissant des renseignements sur le ménage ne possédant qu'un téléphone mobile est de 30,5 %. Le document California Health Interview Survey (2009) contient une discussion détaillée du plan de l'étude, y compris les différences entre le taux de réponse global pour tous les ménages possédant un téléphone mobile et le taux pour les ménages possédant exclusivement un téléphone mobile.

Données de la CHIS 2007 a été effectuée par Westat de la fin de 2007 au début de 2008.

2. Contexte

La plupart des publications traitant des enquêtes à base de sondage double portent sur la théorie statistique relative à l'efficacité du plan de sondage et de l'estimation. Nous résumons certains résultats importants concernant l'échantillonnage, la pondération et l'estimation de la variance, puis nous discutons de l'application de ces méthodes aux enquêtes téléphoniques à base de sondage double.

2.1 Échantillonnage

Soit les deux bases de sondage désignées par A et B . Nous supposons que les échantillons tirés de ces bases, S_A et S_B , sont indépendants. Le domaine des unités qui ne figurent que dans A est désigné par a , le domaine des unités comprises uniquement dans B est désigné par b , et l'intersection contenant les unités chevauchantes est désignée par ab . Dans notre application aux enquêtes téléphoniques, A est la base de numéros de téléphone fixe, B est la base des numéros de téléphone mobile, a est le domaine des ménages n'ayant que des numéros de téléphone fixe, b est le domaine des ménages n'ayant que des numéros de téléphone mobile, et ab est le domaine des ménages ayant les deux types de service téléphonique. De nombreuses caractéristiques importantes des enquêtes à base de sondage double dépendent du traitement réservé aux unités qui pourraient se trouver dans les deux bases de sondage (domaine ab).

Une approche d'enquête à base de sondage double avec présélection vise à rendre $ab = \emptyset$ en éliminant toutes les unités chevauchantes avant l'échantillonnage, après l'échantillonnage mais avant la collecte des données, durant la collecte des données, ou après la collecte des données. Lohr (2009) donne des exemples d'enquêtes à base de sondage double réalisées selon chacune de ces approches.

Brick, Edwards et Lee (2007) et Fleeman (2007) décrivent la présélection dans les enquêtes téléphoniques à base de sondage double. Bien qu'aux États-Unis, les numéros de téléphone puissent être partitionnés selon qu'il s'agit de numéros de téléphone mobile ou de téléphone fixe, cette base de sondage ne permet pas de déterminer si ces numéros correspondent à des ménages n'ayant que des téléphones fixes (a), de ménages n'ayant que des téléphones mobiles (b), ou de ménages ayant les deux types de service (ab). Dans les enquêtes décrites par Brick, Edwards et Lee (2007) et Fleeman (2007), les ménages tirés de la liste de numéros de téléphone mobile (B) ont été écartés durant la collecte des données s'ils déclaraient posséder une ligne fixe. Cette approche de sélection est celle qui a été adoptée pour la CHIS 2007.

Selon une deuxième approche, appelée enquête à base de sondage double avec chevauchement, les unités faisant partie du domaine de chevauchement peuvent être échantillonnées à partir des deux bases de sondage. Dans ce cas,

2.2 Estimation

Les méthodes d'estimation doivent être appliquées pour éviter de produire des estimations biaisées parce que les unités chevauchantes ont de multiples chances de sélection. Steeh (2004), Brick, Brick, Dipko, Presser, Tucker et Yuan (2007), et Kennedy (2007) discutent des enquêtes téléphoniques à base de sondage double avec chevauchement. Dans ces cas, tous les répondants sont interviewés, quelle que soit la base de sondage dans laquelle ils ont été sélectionnés. L'approche avec chevauchement est utilisée dans les enquêtes Pew.

Dans une enquête avec présélection, la production de pondérations pour l'estimation des totaux et des caractéristiques de l'ensemble de la population est simple, du moins en l'absence d'erreur non due à l'échantillonnage. Puisque $ab = \emptyset$ et que l'échantillonnage est indépendant, les unités échantillonnées dans chaque base de sondage se voient attribuer un poids égal à l'inverse de leur probabilité de sélection dans la base de sondage dont elles proviennent. Une estimation globale du total s'obtient en faisant la somme des estimations de domaine pondérées, $y_{scr} = y_A + y_B$, où $y_A = \sum_{i \in S_A} d_i y_i$ et $y_B = \sum_{i \in S_B} d_i y_i$, où d_i est l'inverse de la probabilité de sélection et $\delta_i(b) = 1$ si i est dans le domaine b et 0 autrement. L'estimation de la variance est également simple, puisque les deux bases de sondage sont des strates et que l'on peut appliquer les méthodes d'estimation de la variance appropriées pour les échantillons stratifiés. Dans le cas des enquêtes téléphoniques, les unités provenant de l'échantillon de numéros de téléphone fixe sont pondérées et ajoutées aux unités pondérées provenant de l'échantillon de numéros de téléphone mobile, après avoir attribué un poids nul aux unités tirées de la base de numéros de téléphone mobile qui possèdent un téléphone fixe.

Même en l'absence d'erreurs non dues à l'échantillonnage, la présélection durant la collecte des données a des conséquences. Par exemple, les ménages tirés de B qui sont écartés ne sont pas admissibles à l'interview, ce qui augmente le coût de la collecte des données et la variance des totaux estimés (Kish 1965, chapitre 11). Les unités qui sont écartées devraient aussi être traitées correctement comme des unités échantillonnées dans l'estimation de la variance. Les enquêtes avec chevauchement sont plus complexes, parce que les unités peuvent être échantillonnées dans l'une ou l'autre base de sondage. Une approche d'estimation consiste à combiner les deux estimations de domaine, y_a et y_b , avec une moyenne des estimations pour la population chevauchante produites d'après les bases de sondage distinctes. Si y_{ab}^A et y_{ab}^B sont les estimations pondérées du domaine de chevauchement provenant des bases de sondage A et B , respectivement, un estimateur moyen ou composite

Erreurs non dues à l'échantillonnage dans les enquêtes téléphoniques à base de sondage double

J. Michael Brick, Ismael Flores Cervantes, Sungho Lee et Greg Norman

Résumé

Les enquêtes téléphoniques à base de sondage double deviennent fréquentes aux États-Unis en raison de l'incomplétude de la liste de numéros de téléphone fixe causée par l'adoption progressive du téléphone mobile. Le présent article traite des erreurs non dues à l'échantillonnage dans les enquêtes téléphoniques à base de sondage double. Alors que la plupart des publications sur les bases de sondage doubles ne tiennent pas compte des erreurs non dues à l'échantillonnage, nous constatons que ces dernières peuvent, dans certaines conditions, causer des biais importants dans les enquêtes téléphoniques à base de sondage double. Nous examinons en particulier les biais dus à la non-réponse et à l'erreur de mesure dans ces enquêtes. En vue de réduire le biais résultant de ces erreurs, nous proposons des méthodes d'échantillonnage à base de sondage double et de pondération. Nous montrons que le facteur de composition utilisé pour combiner les estimations provenant de deux bases de sondage joue un rôle important dans la réduction du biais de non-réponse.

Mots clés : Biais de non-réponse ; erreur de mesure ; calage ; répartition de l'échantillon ; composite.

1. Introduction

Les enquêtes téléphoniques à base de sondage double comportant l'échantillonnage de numéros de téléphone fixe et de numéros de téléphone mobile ont pris de l'importance aux États-Unis en vue de réduire le biais de sous-dénombrement dû à l'incomplétude de la liste de numéros de téléphone fixe. Blumberg et Luke (2009) montrent que le pourcentage de ménages ne possédant pas de ligne téléphonique fixe mais dotés d'au moins un téléphone mobile a augmenté de façon spectaculaire ces dernières années, pour atteindre 20 % à la fin de 2008. D'autres pays font également part d'un accroissement considérable du pourcentage de personnes abonnées seulement à la téléphonie mobile (par exemple Kausela, Callegaro et Vehovar 2008 ; Vicente et Reis 2009).

Le présent article s'appuie sur des données provenant de la California Health Interview Survey (CHIS) et de huit enquêtes réalisées pour le compte du Pew Research Center for the People & the Press (enquêtes Pew) en vue d'étudier les effets des erreurs non dues à l'échantillonnage dans les enquêtes téléphoniques à base de sondage double. La CHIS 2007, menée auprès des Californiens adultes, a été réalisée à la fin de 2007. Cette enquête combine un sondage par téléphone fixe classique et un échantillon de présélection de numéros de téléphone mobile, dans lequel seuls les adultes ayant indiqué que le ménage dont ils faisaient partie ne possédait pas de numéro de téléphone fixe ont été interviewés. Les enquêtes Pew sont des sondages nationaux comportant l'interview d'un adulte pour chacun des numéros de téléphone résidentiels provenant de l'échantillon de

téléphones fixes et de celui de téléphones mobiles. Ces enquêtes sont décrites plus en détail plus loin. La réalisation de ces enquêtes téléphoniques à base de sondage double a mis en relief un certain nombre de questions importantes associées à l'effet des erreurs non dues à l'échantillonnage – erreurs qui n'ont pas été examinées complètement dans d'autres études.

À la section suivante, nous passons en revue le plan de sondage, la pondération et les méthodes d'estimation de la variance élaborées pour les enquêtes à base de sondage double, et décrivons CHIS 2007 et les enquêtes téléphoniques Pew à base de sondage double qui sont utilisées tout au long de l'article. À la troisième section, nous discutons de l'erreur non due à l'échantillonnage dans les enquêtes téléphoniques à base de sondage double, et des effets que ces erreurs peuvent avoir sur le biais des estimations. La non-réponse et les erreurs de mesure revêtent une importance particulière dans les enquêtes à base de sondage double. À la quatrième section, nous étudions les méthodes d'échantillonnage et d'estimation qui peuvent être utilisées pour réduire le biais dans les enquêtes téléphoniques à base de sondage double et nous donnons les conditions sous lesquelles ces approches d'échantillonnage et d'estimation peuvent être les plus utiles. Dans cette section, nous proposons trois estimateurs en vue de réduire le biais dû à la non-réponse différentielle dans le domaine de chevauchement des bases de sondage. À la dernière section, nous résumons certaines constatations concernant les enquêtes téléphoniques à base de sondage double et avançons des hypothèses quant à l'applicabilité de ces constatations à d'autres enquêtes à base de sondage double.



The paper used in this publication meets the minimum requirements of American National Standard for Information Sciences – Permanence of Paper for Printed Library Materials, ANSI Z39.48 - 1984.



Le papier utilisé dans la présente publication répond aux exigences minimales de l'«American National Standard for Information Sciences» – «Permanence of Paper for Printed Library Materials», ANSI Z39.48 - 1984.

Techniques d'enquête

Une revue éditée par Statistique Canada
Volume 37, numéro 1, juin 2011
Table des matières

Articles réguliers

J. Michael Brick, Ismael Flores Cervantes, Sunghae Lee et Greg Norman
Erreurs non dues à l'échantillonnage dans les enquêtes téléphoniques à base de sondage double 1

James O. Chipperfield, Glenys R. Bishop et Paul Campbell
Estimation du maximum de vraisemblance pour les tableaux de contingence et la régression
logistique en présence de données incorrectement apparées 17

Yong You et Qian M. Zhou
Estimation sur petits domaines hiérarchique bayésienne sous un modèle spatial avec application
à des données d'enquête sur la santé 31

Hukum Chandra et Ray Chambers
Estimation sur petits domaines sous linéarisation 45

Sophie Baillargeon et Louis-Paul Rivest
Elaboration de plans stratifiés en R à l'aide du programme *stratification* 59

Jae Kwang Kim et Cindy Long Yu
Estimation de la variance par répliques sous échantillonnage à deux phases 73

Stanislav Kolenikov et Gustavo Angeles
Rentabilité des enquêtes avec échantillonnage en grappes répétées 83

Paul Knothnerus
À propos de l'efficacité de l'échantillonnage à probabilité proportionnelle à la taille aléatoire 103

Communications brèves

Eric Lesage
Utilisation des équations estimantes pour réaliser un calage sur des paramètres complexes 111

Autres revues 117

TECHNIQUES D'ENQUÊTE

Une revue éditée par Statistique Canada

Techniques d'enquête est répertoriée dans The ISI Web of Knowledge (Web of science), The Survey Statistician, Statistical Theory and Methods Abstracts et SRM Database of Social Research Methodology, Erasmus University. On peut en trouver les références dans Current Index to Statistics, et Journal Contents in Qualitative Methods. La revue est également citée par SCOPUS sur les bases de données Elsevier Bibliographic Databases.

COMITÉ DE DIRECTION

Président

J. Kovar

Anciens présidents

D. Royce (2006-2009)

G.J. Brackstone (1986-2005)

R. Platek (1975-1986)

H. Mantel

M.A. Hidiroglou

J. Gambino

S. Fortier (Gestionnaire de la production)

Membres

G. Beaudoin

Rédacteur en chef M.A. Hidiroglou, *Statistique Canada*

Rédacteur en

chef délégué H. Mantel, *Statistique Canada*

Rédacteurs associés

J.-F. Beaumont, *Statistique Canada*

J. van den Brakel, *Statistics Netherlands*

J.M. Brick, *Westat Inc.*

P. Cantwell, *U.S. Bureau of the Census*

R. Chambers, *Centre for Statistical and Survey Methodology*

J.L. Eitinge, *U.S. Bureau of Labor Statistics*

W.A. Fuller, *Iowa State University*

J. Gambino, *Statistique Canada*

B. Hultiger, *University of Applied Sciences Northwestern Switzerland*

D. Judkins, *Westat Inc.*

D. Kasprzyk, *NORC at the University of Chicago*

P. Kott, *National Agricultural Statistics Service*

P. Lahiri, *JPSM, University of Maryland*

P. Lavallée, *Statistique Canada*

P. Lynn, *University of Essex*

D.J. Malec, *National Center for Health Statistics*

G. Nathan, *Hebrew University*

J. Opsomer, *Colorado State University*

Rédacteurs adjoints C. Bocci, P. Dick, G. Dubreuil, S. Godbout, D. Haziza, Z. Patak, S. Rubin-Bleuer et Y. You, *Statistique Canada*

POLITIQUE DE RÉDACTION

Techniques d'enquête publie des articles sur les divers aspects des méthodes statistiques qui intéressent un organisme statistique comme, par exemple, les problèmes de conception découlant de contraintes d'ordre pratique, l'utilisation de différentes sources de données et de méthodes de collecte, les erreurs dans les enquêtes, l'évaluation des enquêtes, la recherche sur les méthodes d'enquête, l'analyse des séries chronologiques, la désaisonnalisation, les études démographiques, l'intégration de données statistiques, les méthodes d'estimation et d'analyse de données et le développement de systèmes généralisés. Une importance particulière est accordée à l'élaboration et à l'évaluation de méthodes qui ont été utilisées pour la collecte de données ou appliquées à des données réelles. Tous les articles seront soumis à une critique, mais les auteurs demeurent responsables du contenu de leur texte et les opinions émises dans la revue ne sont pas nécessairement celles du comité de rédaction ni de Statistique Canada.

Présentation de textes pour la revue

Techniques d'enquête est publiée deux fois l'an. Les auteurs désirant faire paraître un article sont invités à le faire parvenir en français ou en anglais en format électronique et préféablement en Word au rédacteur en chef. (rt@statcan.gc.ca, Statistique Canada, 150 Promenade du Pré Tunney, Ottawa, (Ontario), Canada, K1A 0T6). Pour les instructions sur le format, veuillez consulter les directives présentées dans la revue ou sur le site web (www.statcan.gc.ca).

Abonnement

Le prix de la version imprimée de *Techniques d'enquête* (N° 12-001-XPB au catalogue) est de 58 \$ CA par année. Le prix n'inclut pas les taxes de vente canadiennes. Des frais de livraison supplémentaires s'appliquent aux envois à l'extérieur du Canada : États-Unis 12 \$ CA (6 \$ × 2 exemplaires); autres pays, 20 \$ CA (10 \$ × 2 exemplaires). Un prix réduit est offert aux membres de l'American Statistical Association, l'Association Internationale de Statisticiens d'Enquête, l'American Association for Public Opinion Research, la Société Statistique du Canada et l'Association des statisticiens et statisticiens du Québec. Des versions électroniques sont disponibles sur le site internet de Statistique Canada : www.statcan.gc.ca.

Techniques d'enquête

Une revue
éditée

par Statistique Canada

Juin 2011 • Volume 37 • Numéro 1

Publication autorisée par le ministre responsable de Statistique Canada

© Ministre de l'Industrie, 2011

Tous droits réservés. Le produit ne peut être reproduit et/ou transmis à des personnes ou organisations à l'extérieur de l'organisme du détenteur de licence. Des droits raisonnables d'utilisation du contenu de ce produit sont accordés seulement à des fins de recherche personnelle, organisationnelle ou de politique gouvernementale ou à des fins éducatives. Cette permission comprend l'utilisation du contenu dans des analyses et dans la communication des résultats et conclusions de ces analyses, y compris la citation de quantités limitées de renseignements complémentaires extraits du produit. Cette documentation doit servir à des fins non commerciales seulement. Si c'est le cas, la source des données doit être citée comme suit : Source (ou « Adapté de », s'il y a lieu) : Statistique Canada, année de publication, nom du produit, numéro au catalogue, volume et numéro, période de référence et page(s). Autrement, les utilisateurs doivent d'abord demander la permission écrite aux Services d'octroi de licences, Division des services à la clientèle, Statistique Canada, Ottawa, Ontario, Canada K1A 0T6.

Juin 2011

N° 12-001-XPB au catalogue

Périodicité : semestrielle

ISSN 0714-0045

Ottawa



Comment obtenir d'autres renseignements

Pour toute demande de renseignements au sujet de ce produit ou sur l'ensemble des données et des services de Statistique Canada, visiter notre site Web à www.statcan.gc.ca. Vous pouvez également communiquer avec nous par courriel à infostats@statcan.gc.ca ou par téléphone entre 8 h 30 et 16 h 30 du lundi au vendredi aux numéros suivants :

Centre de contact national de Statistique Canada

Numéros sans frais (Canada et États-Unis) :

Service de renseignements
Service national d'appareils de télécommunications pour les malentendants
Télécopieur 1-800-263-1136
1-800-363-7629
1-877-287-4369

Appels locaux ou internationaux :

Service de renseignements
Télécopieur 1-613-951-8116
1-613-951-0581

Programme des services de dépôt

Service de renseignements
Télécopieur 1-800-635-7943
1-800-565-7757

Comment accéder à ce produit ou le commander

Le produit n° 12-001-X au catalogue est disponible gratuitement sous format électronique. Pour obtenir un exemplaire, il suffit de visiter notre site Web à www.statcan.gc.ca et de parcourir par « Ressource clé » > « Publications ».
Ce produit est aussi disponible en version imprimée standard au prix de 30 \$CAN l'exemplaire et de 58 \$CAN pour un abonnement annuel.

Les frais de livraison supplémentaires suivants s'appliquent aux envois à l'extérieur du Canada :

	Exemplaire	Abonnement annuel
États-Unis	6 \$CAN	12 \$CAN
Autres pays	10 \$CAN	20 \$CAN

Les prix ne comprennent pas les taxes sur les ventes.

La version imprimée peut être commandée par les moyens suivants :

- Téléphone (Canada et États-Unis) 1-800-267-6677
- Télécopieur (Canada et États-Unis) 1-877-287-4369
- Courriel infostats@statcan.gc.ca
- Poste
- Finances
Statistique Canada
Immeuble R.-H.-Coats, 6^e étage
150, promenade Tunney's Pasture
Ottawa (Ontario) K1A 0T6
- En personne auprès des agents et librairies autorisés.

Lorsque vous signalez un changement d'adresse, veuillez nous fournir l'ancienne et la nouvelle adresse.

Normes de service à la clientèle

Statistique Canada s'engage à fournir à ses clients des services rapides, fiables et courtois. À cet égard, notre organisme s'est doté de normes de service à la clientèle que les employés observent. Pour obtenir une copie de ces normes de service, veuillez communiquer avec Statistique Canada au numéro sans frais 1-800-263-1136. Les normes de service sont aussi publiées sur le site www.statcan.gc.ca sous « À propos de nous » > « Notre organisme » > « Offrir des services aux Canadiens ».

Techniques d'enquête

N° 12-001-XPB au catalogue

Une revue
éditée
par Statistique Canada

Juin 2011

•

Volume 37

•

Numéro 1



CA1
BS 12
-C001

Survey Methodology

Catalogue No. 12-001-XPB

A journal
published by
Statistics Canada

December 2011

•

Volume 37

•

Number 2



Statistics
Canada

Statistique
Canada

Canada

How to obtain more information

For information about this product or the wide range of services and data available from Statistics Canada, visit our website at www.statcan.gc.ca, e-mail us at infostats@statcan.gc.ca, or telephone us, Monday to Friday from 8:30 a.m. to 4:30 p.m., at the following numbers:

Statistics Canada's National Contact Centre

Toll-free telephone (Canada and United States):

Inquiries line	1-800-263-1136
National telecommunications device for the hearing impaired	1-800-363-7629
Fax line	1-877-287-4369

Local or international calls:

Inquiries line	1-613-951-8116
Fax line	1-613-951-0581

Depository Services Program

Inquiries line	1-800-635-7943
Fax line	1-800-565-7757

To access and order this product

This product, Catalogue no. 12-000-X, is available free in electronic format. To obtain a single issue, visit our website at www.statcan.gc.ca and browse by "Key resource" > "Publications."

This product is also available as a standard printed publication at a price of CAN\$30.00 per issue and CAN\$58.00 for a one-year subscription.

The following additional shipping charges apply for delivery outside Canada:

	Single issue	Annual subscription
United States	CAN\$6.00	CAN\$12.00
Other countries	CAN\$10.00	CAN\$20.00

All prices exclude sales taxes.


The printed version of this publication can be ordered as follows:

- Telephone (Canada and United States) 1-800-267-6677
- Fax (Canada and United States) 1-877-287-4369
- E-mail infostats@statcan.gc.ca
- Mail
Statistics Canada
Finance
R.H. Coats Bldg., 6th Floor
150 Tunney's Pasture Driveway
Ottawa, Ontario K1A 0T6
- In person from authorized agents and bookstores.

When notifying us of a change in your address, please provide both old and new addresses.

Standards of service to the public

Statistics Canada is committed to serving its clients in a prompt, reliable and courteous manner. To this end, Statistics Canada has developed standards of service that its employees observe. To obtain a copy of these service standards, please contact Statistics Canada toll-free at 1-800-263-1136. The service standards are also published on www.statcan.gc.ca under "About us" > "The agency" > "Providing services to Canadians."



Survey Methodology

A journal
published by
Statistics Canada

December 2011 • Volume 37 • Number 2

Published by authority of the Minister responsible for Statistics Canada

© Minister of Industry, 2011

All rights reserved. This product cannot be reproduced and/or transmitted to any person or organization outside of the licensee's organization. Reasonable rights of use of the content of this product are granted solely for personal, corporate or public policy research, or for educational purposes.

This permission includes the use of the content in analyses and the reporting of results and conclusions, including the citation of limited amounts of supporting data extracted from this product. These materials are solely for non-commercial purposes. In such cases, the source of the data must be acknowledged as follows: Source (or "Adapted from", if appropriate): Statistics Canada, year of publication, name of product, catalogue number, volume and issue numbers, reference period and page(s). Otherwise, users shall seek prior written permission of Licensing Services, Client Services Division, Statistics Canada, Ottawa, Ontario, Canada K1A 0T6.

December 2011

Catalogue no. 12-001-XPB

Frequency: Semi-annual

ISSN 0714-0045

Ottawa



Statistics Canada
Statistique Canada

Canada

SURVEY METHODOLOGY

A Journal Published by Statistics Canada

Survey Methodology is indexed in The ISI Web of knowledge (Web of science), The Survey Statistician, Statistical Theory and Methods Abstracts and SRM Database of Social Research Methodology, Erasmus University and is referenced in the Current Index to Statistics, and Journal Contents in Qualitative Methods. It is also covered by SCOPUS in the Elsevier Bibliographic Databases.

MANAGEMENT BOARD

Chairman J. Kovar

Past Chairmen D. Royce (2006-2009)
G.J. Brackstone (1986-2005)
R. Platek (1975-1986)

Members G. Beaudoin
S. Fortier (Production Manager)
J. Gambino
M.A. Hidiroglou
H. Mantel

EDITORIAL BOARD

Editor M.A. Hidiroglou, *Statistics Canada*
Deputy Editor H. Mantel, *Statistics Canada*

Past Editor J. Kovar (2006-2009)
M.P. Singh (1975-2005)

Associate Editors

J.-F. Beaumont, *Statistics Canada*
J. van den Brakel, *Statistics Netherlands*
J.M. Brick, *Westat Inc.*
P. Cantwell, *U.S. Bureau of the Census*
R. Chambers, *Centre for Statistical and Survey Methodology*
J.L. Eltinge, *U.S. Bureau of Labor Statistics*
W.A. Fuller, *Iowa State University*
J. Gambino, *Statistics Canada*
D. Haziza, *Université de Montréal*
B. Hulliger, *University of Applied Sciences Northwestern Switzerland*
D. Judkins, *Westat Inc.*
D. Kasprzyk, *NORC at the University of Chicago*
P. Kott, *RTI International*
P. Lahiri, *JPSM, University of Maryland*
P. Lavallée, *Statistics Canada*
P. Lynn, *University of Essex*
D.J. Malec, *National Center for Health Statistics*
G. Nathan, *Hebrew University*
J. Opsomer, *Colorado State University*

D. Pfeffermann, *Hebrew University*
N.G.N. Prasad, *University of Alberta*
J.N.K. Rao, *Carleton University*
J. Reiter, *Duke University*
L.-P. Rivest, *Université Laval*
N. Schenker, *National Center for Health Statistics*
F.J. Scheuren, *National Opinion Research Center*
P. do N. Silva, *Escola Nacional de Ciências Estatísticas*
P. Smith, *Office for National Statistics*
E. Stasny, *Ohio State University*
D. Steel, *University of Wollongong*
L. Stokes, *Southern Methodist University*
M. Thompson, *University of Waterloo*
V.J. Verma, *Università degli Studi di Siena*
K.M. Wolter, *National Opinion Research Center*
C. Wu, *University of Waterloo*
W. Yung, *Statistics Canada*
A. Zaslavsky, *Harvard Medical School*

Assistant Editors C. Bocci, K. Bosa, P. Dick, G. Dubreuil, S. Godbout, Z. Patak, S. Rubin-Bleuer and
Y. You, *Statistics Canada*

EDITORIAL POLICY

Survey Methodology publishes articles dealing with various aspects of statistical development relevant to a statistical agency, such as design issues in the context of practical constraints, use of different data sources and collection techniques, total survey error, survey evaluation, research in survey methodology, time series analysis, seasonal adjustment, demographic studies, data integration, estimation and data analysis methods, and general survey systems development. The emphasis is placed on the development and evaluation of specific methodologies as applied to data collection or the data themselves. All papers will be refereed. However, the authors retain full responsibility for the contents of their papers and opinions expressed are not necessarily those of the Editorial Board or of Statistics Canada.

Submission of Manuscripts

Survey Methodology is published twice a year. Authors are invited to submit their articles in English or French in electronic form, preferably in Word to the Editor, (smj@statcan.gc.ca, Statistics Canada, 150 Tunney's Pasture Driveway, Ottawa, Ontario, Canada, K1A 0T6). For formatting instructions, please see the guidelines provided in the journal and on the web site (www.statcan.gc.ca).

Subscription Rates

The price of printed versions of *Survey Methodology* (Catalogue No. 12-001-XPB) is CDN \$58 per year. The price excludes Canadian sales taxes. Additional shipping charges apply for delivery outside Canada: United States, CDN \$12 (\$6 × 2 issues); Other Countries, CDN \$20 (\$10 × 2 issues). A reduced price is available to members of the American Statistical Association, the International Association of Survey Statisticians, the American Association for Public Opinion Research, the Statistical Society of Canada and l'Association des statisticiennes et statisticiens du Québec. Electronic versions are available on Statistics Canada's web site: www.statcan.gc.ca.

Survey Methodology
A Journal Published by Statistics Canada
Volume 37, Number 2, December 2011

Contents

Waksberg Invited Paper Series

Danny Pfeffermann Modelling of complex survey data: Why model? Why is it a problem? How can we approach it?	115
--	-----

Regular Papers

Balgobin Nandram and Hasanjan Sayit A Bayesian analysis of small area probabilities under a constraint.....	137
Ray Chambers, Hukum Chandra and Nikos Tzavidis On bias-robust mean squared error estimation for pseudo-linear small area estimators	153
Jean-François Beaumont and Joël Bissonnette Variance estimation under composite imputation: The methodology behind SEVANI	171

Special Section of the U.S. Census Bureau

Introduction

Patrick E. Flanagan and Ruth Ann Killion Alternative demographic sample designs being explored at the U.S. Census Bureau.....	181
--	-----

Special Section Papers

Steve Thompson Adaptive network and spatial sampling	183
Sharon L. Lohr Alternative survey sample designs: Sampling with multiple overlapping frames	197
Yves Tillé Ten years of balanced sampling with the cube method: An appraisal	215

Discussion

Jean Opsomer Innovations in survey sampling design: Discussion of three contributions presented at the U.S. Census Bureau....	227
--	-----

Acknowledgements	233
-------------------------------	-----

Announcements	235
----------------------------	-----

In Other Journals	237
--------------------------------	-----

The paper used in this publication meets the minimum requirements of American National Standard for Information Sciences – Permanence of Paper for Printed Library Materials, ANSI Z39.48 - 1984.



Le papier utilisé dans la présente publication répond aux exigences minimales de l'American National Standard for Information Sciences – “Permanence of Paper for Printed Library Materials”, ANSI Z39.48 - 1984.



Waksberg Invited Paper Series

The journal *Survey Methodology* has established an annual invited paper series in honour of Joseph Waksberg, who has made many important contributions to survey methodology. Each year a prominent survey researcher is chosen to author an article as part of the Waksberg Invited Paper Series. The paper reviews the development and current state of a significant topic within the field of survey methodology, and reflects the mixture of theory and practice that characterized Waksberg's work.

Please see the announcements at the end of the Journal for information about the nomination and selection process of the 2012 Waksberg Award.

This issue of *Survey Methodology* opens with the tenth paper of the Waksberg Invited Paper Series. The editorial board would like to thank the members of the selection committee Daniel Kasprzyk (Chair), Elisabeth A. Martin, Mary E. Thompson and Wayne Fuller for having selected Danny Pfeffermann as the author of this year's Waksberg Award paper.

2011 Waksberg Invited Paper

Author: Danny Pfeffermann

Danny Pfeffermann is Professor of statistics at the Hebrew University of Jerusalem, Israel, and at Southampton Statistical Sciences Research Institute (S3RI), University of Southampton, UK. For the past 15 years he is also a consultant for the US Bureau of Labor Statistics. His main research areas are analytic inference from complex sample surveys, seasonal adjustment and trend estimation, small area estimation, and more recently, observational studies and nonresponse. Danny served for two years as the president of the Israel Statistical Association and is the president elect of the International Association of Survey Statisticians (IASS). He is co-editor of the new two-volume handbook in Statistics on "Sample Surveys".

Waksberg Invited Paper Series**Preface from the author**

It is a great honour to receive the award named after Joe Waksberg. I am old enough to have had the fortune of meeting Joe on several occasions, the last time being a whole day of professional meetings at Westat, discussing nothing else but my own modest contributions to survey sampling. What I remember from these meetings is Joe's brilliance, profound knowledge and sharp intellect, even at his very advanced age. I would be lying if I say that I was able to answer all his critical questions.

I feel even more honoured and privileged when I look at the list of all the eminent survey statisticians who received the award before me. While I am still trying to convince myself that I deserve being on that list, I am overwhelmed by all the sincere congratulations and good words from colleagues around the world and during the symposium. What can I say, I am very proud and grateful.

On this occasion, I would like to commemorate also one of the founders and the long serving editor of *Survey Methodology*, the late M.P. Singh. In 1993 I published a paper in the *International Statistical Review* entitled "The role of sampling weights when modeling survey data". This paper was well received and when I met M.P. a couple of years later, he sort of complained to me for not publishing the paper in *Survey Methodology*. Not having a convincing answer, I promised M.P. that one day I would write another paper on this topic and submit it to *Survey Methodology*. I feel that with the present paper I have kept my promise to M.P. Singh.

Danny Pfeffermann

Modelling of complex survey data: Why model? Why is it a problem? How can we approach it?

Danny Pfeffermann¹

Abstract

This article attempts to answer the three questions appearing in the title. It starts by discussing unique features of complex survey data not shared by other data sets, which require special attention but suggest a large variety of diverse inference procedures. Next a large number of different approaches proposed in the literature for handling these features are reviewed with discussion on their merits and limitations. The approaches differ in the conditions underlying their use, additional data required for their application, goodness of fit testing, the inference objectives that they accommodate, statistical efficiency, computational demands, and the skills required from analysts fitting the model. The last part of the paper presents simulation results, which compare the approaches when estimating linear regression coefficients from a stratified sample in terms of bias, variance, and coverage rates. It concludes with a short discussion of pending issues.

Key Words: Informative sampling; NMAR nonresponse; Likelihood-based methods; Probability weighting; Randomization distribution; Sample model.

1. Introduction

Survey data are frequently used for analytic inference on statistical models, which are assumed to hold for the population from which the sample is taken. Familiar examples include the estimation of income elasticities from household surveys, the analysis of labour market dynamics from labour force surveys, comparisons of pupils' achievements from educational surveys and the search for causal relationships between risk factors and disease prevalence from health surveys. An important common feature to all these examples is that interest lies in the structure of the models being estimated and what can be learnt from them. This is different from fitting models merely for prediction purposes, such as when predicting finite population totals or in small area estimation, where the structure and interpretation of the model are of secondary importance. Models are also used implicitly for choosing the sampling design and estimators, such as in stratified sampling, or when defining weighting cells for nonresponse adjustments. However, inference is typically based in these cases on the randomization distribution over all possible sample selections, and not on the model, which is known as 'model assisted inference'.

Survey data typically differ from other data sets in five main aspects.

1. The samples are selected at random with known selection probabilities, which allows using the randomization distribution over all possible sample selections as the basis for inference instead of the hypothetical distribution underlying the population model. As discussed below, a combination of the two distributions is in common use.

2. The sample selection probabilities in at least some stages of the sample selection are often unequal; when these probabilities are related to the model outcome variable, the sampling process becomes informative and the model holding for the sample is then different from the target population model.
3. Survey data are almost inevitably subject to various forms of nonresponse, often of considerable magnitude, which again may distort the population model if the response propensity is associated with the outcome of interest (not missing at random non-response).
4. The sample data are often clustered due to the use of multi-stage cluster samples. The clusters are 'natural units' (households, individuals in case of longitudinal surveys...), implying that observations within the same cluster are correlated.
5. The data available to the modeler may be masked ("swapped", "contaminated", suppressed") in order to protect the anonymity of the respondents. When this is the case, the modeler's data differ from the correct data.

Many approaches have been proposed in the literature for estimating population models from complex survey data possessing these features, some of which are more familiar than the others. The approaches differ in the conditions underlying their use, the data required for their application, goodness of fit testing, the inference objectives that they accommodate, statistical efficiency, computational demands, and the skills required from analysts fitting the model. This heterogeneity means that there does not exist any single

1. Danny Pfeffermann, Southampton Statistical Sciences Research Institute, U.K. and Hebrew University of Jerusalem, Israel. E-mail: d.pfeffermann@soton.ac.uk.

approach that can be considered as best in all situations. That being the case, a fundamental question arising is which approach or approaches could or should be used for a given practical application.

The present paper is divided into three parts. In the first part (Section 2) I elaborate on the first four features of complex survey data mentioned above. In the second part (Section 3) I review the various approaches proposed in the literature for dealing with these features, discussing their merits and limitations in light of the properties mentioned above. In the third part (Section 4) I present simulation results which compare the approaches when estimating a linear regression model from a stratified sample in terms of bias, variance, and coverage rates. I conclude with a short discussion of pending issues in Section 5.

2. Why are survey data different from other data?

2.1 The problem of unequal sampling probabilities and nonresponse

Consider a finite population $U = \{1, \dots, N\}$ with measurements $\{y_i, x_i, z_i\}$ for unit $i = 1, \dots, N$, where y represents an outcome variable of interest, x a vector of covariates and z a vector of design variables used for the sample selection. The design variables may include some or all of the covariates, and in special cases also the outcome variable when known for all the population units, such as in case-control studies. The matrix $Z_U = [z_1, \dots, z_N]$ is known to the sampler drawing the sample, but not necessarily to the analyst fitting the model. Denote by $s = (I_1, \dots, I_N)$ the selected sample, where I_i is the sampling indicator taking the value 1 if unit $i \in U$ is drawn to the sample and 0 otherwise. In practice, not all the sampled units necessarily respond, and we denote by R_i the response indicator; $R_i = 1(0)$ if unit $i \in S$ responds (does not respond).

The observed data may be viewed as the outcome of three random processes. The first process generates the vectors $\{y_i, x_i, z_i\}$ for the N population units. The second process selects a sample s from U at random by a sampling design, $\Pr(s) = \Pr(s | Z_U)$. The third process selects the responding units. This process is obviously not part of the original sampling design and is often the result of 'self selection', although nonresponse could be caused by many other reasons. See Brick and Montaquila (2009) for a recent overview.

When the sample selection probabilities and/or the response probabilities are related to the values of the outcome variable even after conditioning on the model covariates, in the sense that $\Pr(I_i = 1 | y_i, x_i) \neq \Pr(I_i = 1 | x_i)$ or $\Pr(R_i = 1 | y_i, x_i, I_i = 1) \neq \Pr(R_i | x_i, I_i = 1)$, the model holding for the observed outcomes is different from the population model. In symbols, $f_o(y_i | x_i) \neq f_p(y_i | x_i)$, where $f_o(y_i | x_i)$

represents the model holding for a unit selected to the sample and responding, and $f_p(y_i | x_i)$ is the *population* model (the model holding for the population values). See Equations (2.1) and (2.2) below.

Example 1. Suppose that the population model is the regression model, $f_p(y_i | x_i) = N(x_i' \beta, \sigma_e^2)$, and that the sample is selected with selection probabilities satisfying $\Pr(I_i = 1 | y_i, x_i) = \exp[\gamma_1 y_i + \gamma_2 y_i^2 + g(x_i)]$, where γ_1 and $\gamma_2 \leq 0$ are constants and $g(x_i)$ is some nonstochastic function of the covariates. Simple use of Bayes theorem (see below) shows that the model holding for the sample outcomes is in this case, $f_s(y_i | x_i) = N[(\gamma_1 \sigma_e^2 + x_i' \beta) / C, \sigma_e^2 / C]$, where $C = (1 - 2\sigma_e^2 \gamma_2)$. Thus, although the sample residuals have again a normal distribution, the regression coefficients and the residual variance are different from their values under the population model. In the special case $\gamma_2 = 0$, the slope coefficients and the residual variance are the same as under the population model, but not the intercept. If $\gamma_1 = 0$ as well, the sample selection probabilities satisfy $\Pr(I_i = 1 | y_i, x_i) = \Pr(I_i = 1 | x_i)$ and the two models are now the same.

Following conventional terminology, when $\Pr(I_i = 1 | y_i, x_i) \neq \Pr(I_i = 1 | x_i)$ the sampling design is said to be *informative*. When $\Pr(R_i = 1 | y_i, x_i, I_i = 1) \neq \Pr(R_i | x_i, I_i = 1)$, the nonresponse is *not missing at random* (NMAR nonresponse). Notice that whereas the sampling probabilities are typically known to the analyst fitting the model, at least for the sampled units, the response probabilities are generally unknown and need to be modelled under NMAR nonresponse. Ignoring an informative sample or NMAR nonresponse and thus assuming implicitly that the model holding for the observed outcomes is the same as the target population model may yield large biases and erroneous inference. The books edited by Kasprzyk, Duncan, Kalton and Singh (1989), Skinner, Holt and Smith (1989) and Chambers and Skinner (2003) contain many discussions and illustrations of the effect of ignoring informative sampling or NMAR nonresponse. See also Pfeffermann (1993, 1996), Pfeffermann and Sverchkov (2009) and Pfeffermann and Sikov (2011) for further discussions and examples, with many other more recent references.

In what follows, I use the abbreviation "*pdf*" to define the probability density function when the outcome is continuous or the probability function when the outcome is discrete. Suppose first that there is no nonresponse. Following Pfeffermann, Krieger and Rinott (1998a), the *marginal sample pdf*, $f_s(y_i | x_i)$ defines the conditional *pdf* of y_i given that unit i is in the sample ($I_i = 1$). By Bayes theorem,

$$\begin{aligned} f_s(y_i | x_i) &= f(y_i | x_i, I_i = 1) \\ &= \frac{\Pr(I_i = 1 | x_i, y_i) f_p(y_i | x_i)}{\Pr(I_i = 1 | x_i)}, \quad (2.1) \end{aligned}$$

where $f_p(y_i | x_i)$ is the corresponding population *pdf*. The probabilities $\Pr(I_i = 1 | x_i, y_i)$ are generally not the same as the sample selection probabilities $\pi_i = \Pr(I_i = 1)$, which may depend on all the population values Z_U of the design variables. However, the use of the marginal sample *pdf* only requires modelling $\Pr(I_i = 1 | x_i, y_i)$. Typically, $\Pr(I_i = 1 | \pi_i, y_i, x_i) = \pi_i$, in which case $\Pr(I_i = 1 | y_i, x_i) = E_p(\pi_i | y_i, x_i)$, where $E_p(\cdot)$ is the expectation under the population *pdf*.

Remark 1. In practice, the covariates featuring in the population model need not be the same as the covariates featuring in the model of the conditional sample inclusion probabilities, $\Pr(I_i = 1 | x_i, y_i)$. In fact, following the results in Pfeffermann and Landsman (2011), identifiability of the sample model often requires that the two sets of covariates are not identical. However, to simplify the presentation in this paper, I assume for convenience that the covariates contained in the population model and the covariates defining the conditional inclusion probabilities are the same, or alternatively, that x_i defines the union of the two sets of covariates.

It follows from (2.1) that unless $\Pr(I_i = 1 | x_i, y_i) = \Pr(I_i = 1 | x_i) \forall y_i$, the sample *pdf* is different from the population *pdf*, in which case the sampling design is informative and cannot be ignored in the inference process. In particular, it follows from (2.1) that under informative sampling,

$$E_s(y_i | x_i) = E_p \left[\frac{\Pr(I_i = 1 | x_i, y_i) y_i}{\Pr(I_i = 1 | x_i)} \middle| x_i \right] \neq E_p(y_i | x_i),$$

where $E_s(\cdot)$ is the expectation under the sample *pdf*. Estimating $E_p(y_i | x_i)$ is often the main target of inference, illustrating that ignoring an informative sampling scheme and thus estimating implicitly $E_s(y_i | x_i)$ can bias the inference.

Suppose now the existence of NMAR nonresponse. The marginal sample *pdf* (2.1) can be extended to this case by defining,

$$\begin{aligned} f_o(y_i | x_i) &= f(y_i | x_i, I_i = 1, R_i = 1) \\ &= \frac{\Pr(R_i = 1 | y_i, x_i, I_i = 1) \Pr(I_i = 1 | y_i, x_i) f_p(y_i | x_i)}{\Pr(R_i = 1 | x_i, I_i = 1) \Pr(I_i = 1 | x_i)} \\ &= \frac{\Pr(R_i = 1 | y_i, x_i, I_i = 1) f_s(y_i | x_i)}{\Pr(R_i = 1 | x_i, I_i = 1)}. \end{aligned} \quad (2.2)$$

Notice from (2.2) that unless $\Pr(R_i = 1 | y_i, x_i, I_i = 1) = \Pr(R_i = 1 | x_i, I_i = 1) \forall y_i$, the *pdf* holding for the observed outcomes is different from the sample *pdf*. Here again I assume for convenience that the response probabilities depend on the same covariates as in the sample model. See Remark 1 above.

The *pdfs* (2.1) and (2.2) define the marginal distributions of the outcome for a given unit. These definitions generalize very naturally to the joint *pdf* of two or more outcomes associated with different units. More generally, define for every plausible sample $s \subset U$ the sample indicator A_s , such that $A_s = 1$ if s is sampled and $A_s = 0$ otherwise, and assume for convenience full response. Denote the data associated with s by (y_s, x_s) . The joint sample *pdf* of $y_s | x_s$ is then,

$$\begin{aligned} f_s(y_s | x_s) &= f(y_s | x_s, A_s = 1) \\ &= \frac{\Pr(A_s = 1 | y_s, x_s) f_p(y_s | x_s)}{\Pr(A_s = 1 | x_s)}. \end{aligned} \quad (2.3)$$

The *pdf* $f_p(y_s | x_s)$ can be general, allowing in particular for correlated measurements, but modelling the probability $\Pr(A_s = 1 | y_s, x_s)$ is practically only feasible if the sample can be decomposed into exclusive and exhaustive subsets s_k such that $\Pr(A_s = 1 | y_s, x_s) \propto \prod_k \Pr(A_{s_k} = 1 | y_{s_k}, x_{s_k})$ and $\Pr(A_{s_k} = 1 | y_{s_k}, x_{s_k})$ satisfies the same model for all the subsets (see Example 2). In particular, if the population outcomes are independent given the covariates under the population model and $\Pr(A_s = 1 | y_s, x_s) \propto \prod_{i \in s} \Pr(I_i = 1 | y_i, x_i)$, (2.3) takes the form

$$\begin{aligned} f_s(y_s | x_s) &= \prod_{i \in s} \frac{\Pr(I_i = 1 | y_i, x_i) f_p(y_i | x_i)}{\Pr(I_i = 1 | x_i)} \\ &= \prod_{i \in s} f_s(y_i | x_i), \end{aligned} \quad (2.4)$$

so that the sample outcomes are likewise independent.

Example 2. Consider the case of a clustered population $U = \bigcup_l U_l$, with independent measurements between clusters, such that $f_p(y_U | x_U) = \prod_l f_p(y_{U_l} | x_{U_l})$, where (y_U, x_U) defines all the population values and (y_{U_l}, x_{U_l}) the values in cluster l . Let s define the set of sampled clusters, assumed to be drawn independently with probabilities $\Pr(l \in s | y_{U_l}, x_{U_l}) = r(y_{U_l}, x_{U_l})$ for some function $r(\cdot)$, and suppose also that all the units in the sampled clusters are observed (single-stage cluster sampling). Then, $\Pr(A_s = 1 | y_U, x_U) = \prod_{k \in s} r(y_{U_k}, x_{U_k}) \times \prod_{j \notin s} [1 - r(y_{U_j}, x_{U_j})]$. Since for $k \in s, (y_{U_k}, x_{U_k}) = (y_{s_k}, x_{s_k})$, it follows that $\Pr(A_s = 1 | y_s, x_s) = \prod_{k \in s} r(y_{s_k}, x_{s_k}) \times G$, where for given covariates $x_{U_j}, j \notin s$, G is a constant satisfying, $G = \prod_{j \notin s} [1 - r(y_{U_j}, x_{U_j})] f_p(y_{U_j} | x_{U_j}) dy_{U_j}$. The case of a non-clustered population with independent measurements and Poisson sampling of individual units is a special case where each cluster consists of a single element, giving rise to (2.4).

Remark 2. The examples considered so far assume independent sampling, which preserves the independence of the outcomes after sampling, but this assumption can

usually be relaxed following a result proved and illustrated in Pfeffermann *et al.* (1998a). By this result, under some general regularity conditions and for many commonly used sampling schemes for selection with unequal probabilities, if the population measurements are independent, the sample measurements are *asymptotically independent* under the sample distribution. The asymptotic framework requires that the population size increases but the sample size is held fixed. As illustrated in section 2.3, the assumption of independent population measurements is often also not restrictive.

So far, we suppressed for convenience from the notation the parameters underlying the population *pdf* and the sampling process. Consider, for example, the sample *pdf* (2.3). With added parameter notation, it can be written as

$$f_s(y_s | x_s; \theta, \gamma) = \frac{\Pr(A_s = 1 | y_s, x_s; \gamma) f_p(y_s | x_s; \theta)}{\Pr(A_s = 1 | x_s; \theta, \gamma)}. \quad (2.5)$$

Thus, the conditional population and sample *pdfs* are different, unless

$$\Pr(A_s = 1 | y_s, x_s; \gamma) = \Pr(A_s = 1 | x_s; \theta, \gamma) \quad \forall y_s. \quad (2.6)$$

When (2.6) holds, inference on the target parameter θ can be implemented by fitting the population model to the sample data, ignoring the sample selection. Note that this conclusion refers to the selected sample defined by the event $A_s = 1$.

The condition (2.6) is a strong condition. In a fundamental article on missing values, Rubin (1976) establishes conditions under which the sampling process can be ignored for likelihood, Bayesian or sampling theory (repeated sampling from a model) inference, that is, conditions under which the population model defined by $f_p(y_s | x_s; \theta)$ can be fitted to the observed data, depending on the inference method used. Little (1982) extends Rubin's results by distinguishing between the sample selection and the response process. Another important distinction is that Little conditions on the population values Z_U of the design variables used for the sample selection. Inference on the target population model $f_p(y_s | x_s; \theta)$ requires therefore integrating the conditional *pdf* of $y_s | Z_U, x_s$ over the distribution of $Z_U | x_s$ (see Section 3). Sugden and Smith (1984) establish conditions under which a sampling process that depends on design variables Z is ignorable, given partial information on the design. Let $d_s = D_s(z_U)$ contain all the available design information for a sample s such as strata membership (may only be known for the sampled units), sample selection probabilities *etc.* Using previous notation, a key condition for ignorability of the sampling process given the available design information is that $A_s \perp Z_U | d_s$, with " \perp " meaning independence, implying $\Pr(A_s = 1 | Z_U = z_U) = \Pr(A_s = 1 | d_s)$ for any z_U for which $D_s(z_U) = d_s$.

For large scale multi-stage sample surveys with possibly many design variables, it is generally difficult and often impractical to check directly the conditions that permit ignoring the sample selection or nonresponse given the available design information. On the other hand, even when the sample *pdf* is different from the population *pdf*, it does not necessarily imply that inference that ignores the sampling process is wrong. As a simple illustration, consider the special case of Example 1 where $\gamma_2 = 0$. In this case the sample *pdf* is normal with the same slope coefficients and residual variance as under the population *pdf*. Thus, for inference about the slope coefficients one can ignore the sampling process. A similar result holds for logistic models when the sample selection depends on y but not on x . See Pfeffermann *et al.* (1998a) for derivation of this result. Pfeffermann and Sverchkov (2009) review several test statistics proposed in the literature for assessing whether ignoring the sample selection is justified for the intended inference.

2.2 The use of the randomization distribution for inference

A unique feature of sample surveys is that the sample is selected at random by use of a sampling design $[\{s, \Pr(s)\}, s \in S]$. The sampling design induces a (discrete) *randomization distribution* for any statistic T_{ys} , which is the conditional distribution over all possible sample selections, given the finite population values. Thus, the statistic T_{ys} takes the value t_{ys} with probability $\Pr(s)$, $s \in S$. Classical survey sampling inference is based solely on this distribution. For example, the familiar Horvitz-Thompson (HT) estimator T_{ys}^{HT} , which takes the value $t_{ys}^{HT} = \sum_{i \in s} (y_i / \pi_i)$ if sample s is drawn, is randomization-unbiased for the finite population total $TOT_y = \sum_{j=1}^N y_j$, since $\sum_{s \in S} \Pr(s) t_{ys}^{HT} = T_y$. Its variance is, $\text{Var}(T_{ys}^{HT}) = \sum_{s \in S} \Pr(s) (t_{ys}^{HT} - T_y)^2$. Notice that in the case of nonresponse, the use of the randomization distribution requires knowledge of the response probabilities, which in practice can only be estimated. The HT estimator takes in this case the form, $T_{ys}^{HT} = \sum_{i \in R} y_i / [\pi_i \times \hat{\Pr}(R_i = 1 | I_i = 1)]$, where R defines the subsample of respondents. See Fuller (2002) for further discussion.

The randomization distribution conditions on the realized population values. Consequently, it can be used for descriptive inference on known functions of the finite population values, but not for analytic inference on a hypothesized model giving rise to these values. For this, one may consider the joint distribution over all possible sample outcomes for given population values (the *randomization r-distribution*) and all possible realizations of the finite population measurements (the *model p-distribution*). See Binder and Roberts (2009) and the references therein. The combined *r-p* distribution offers an alternative framework of

inference to the use of the *pdfs* $f_s(y | x)$ or $f_o(y | x)$ defined before.

Example 3: Suppose that the population model is $y_i \sim \text{Mult}[\{p_k\}, K]$, such that $\Pr_p(y_i = k) = p_k$, $k = 1, \dots, K$; $\sum_{k=1}^K p_k = 1$. Let $\Pr(i \in s | y_i = k) = \pi_k$. Then, by (2.1), $\Pr_s(y_i = k) = \Pr(y_i = k | i \in s) = \pi_k p_k / \sum_{j=1}^K \pi_j p_j = p_k^*$, or, $y_i | i \in s \sim \text{Mult}(\{p_k^*\}, K)$. Assuming independence of the observed outcomes and known selection probabilities, the maximum likelihood estimator (*mle*) of p_k based on the sample distribution is $\tilde{p}_k = (n_k / \pi_k) / \sum_{j=1}^K (n_j / \pi_j)$, where n_k is the number of sampled units with outcome $y_i = k$. The use of the $r-p$ distribution suggests estimating p_k by the HT estimator $\hat{p}_k = (1/N) \sum_{i|y_i=k} (1/\pi_k) = (n_k / \pi_k) / N$. The estimator \hat{p}_k is randomization-unbiased for $\hat{P}_k = N_k / N$, where N_k is the number of population units with outcome $y_j = k$, and \hat{P}_k is p -unbiased for p_k , such that \hat{p}_k is $r-p$ -unbiased for p_k .

The obvious difference between the $r-p$ distribution and the sample distribution, $f_s(y | x)$, is that the latter conditions on the observed sample of units (and hence the observed values of the covariates or the selected clusters in a cluster sample), whereas the $r-p$ distribution accounts for all possible sample selections. Consequently, the use of the latter distribution does not lend itself in general to conditional inference. The use of the *pdfs* $f_s(y | x)$ or $f_o(y | x)$ requires modelling $\Pr(I_i = 1 | x_i, y_i)$ (Equation 2.1) and $\Pr(R_i = 1 | y_i, x_i, I_i = 1)$ in case of nonresponse (Equation 2.2), but it permits the computation (estimation) of the conditional *pdf* of the observed outcomes given the covariates, and hence the use of classical inference tools.

2.3 Data obtained from a cluster sample

Another special feature of survey data mentioned in the introduction is *clustering*, due to the use of multi-stage cluster samples. The clusters are 'natural groups' such as households, residence blocks, schools, or even individuals in the case of longitudinal surveys. Consequently, the outcomes pertaining to the same cluster are generally correlated, known as the *intraclass correlation*. It is important to emphasize that the clusters represent an existing population grouping, such that an intraclass correlation exists also under the population model.

Pfeffermann and Smith (1985) review several classes of plausible regression models for clustered populations, and discuss how they can be estimated from the sample. A population model in common use is the random intercept model,

$$y_{ij} = \underset{\text{indep.}}{x'_{ij}} \beta + u_i + \underset{\text{indep.}}{\varepsilon_{ij}}, \quad i = 1, \dots, M, \quad j = 1, \dots, N_i; \\ u_i \sim N(0, \sigma_u^2); \quad \varepsilon_{ij} \sim N(0, \sigma_\varepsilon^2), \quad (2.7)$$

where M defines the number of clusters and N_i the number of units in cluster i . The model assumes also $E(u_i \varepsilon_{ij}) = 0, \forall i, j$. Notice that under this model $\text{Var}(y_{ij}) = \sigma_u^2 + \sigma_\varepsilon^2$,

$E(y_{ij} y_{il}) = \sigma_u^2$ for $j \neq l$ and $E(y_{ij} y_{kl}) = 0$ for $i \neq k$, implying

$$\begin{aligned} \text{Corr}(y_{ij}, y_{il}) &= \sigma_u^2 / (\sigma_u^2 + \sigma_\varepsilon^2) \quad \text{for } j \neq l; \\ \text{Corr}(y_{ij}, y_{kl}) &= 0 \quad \text{for } i \neq k. \end{aligned} \quad (2.8)$$

Scott and Holt (1982) show that estimating β in (2.7) by ordinary least squares (OLS) usually results in a small loss of efficiency, compared to the use of the optimal generalized least squares (GLS) estimator. However, ignoring the intra-cluster correlation when estimating the variance of the OLS estimator may result in considerable variance underestimation and hence wrong size and excessive powers of test statistics and too short confidence intervals.

The results in Scott and Holt (1982) and Pfeffermann and Smith (1985) assume noninformative sampling and full response. When this is not the case, the model holding for the sample data is different from the corresponding population model, although the clustered nature of the model is preserved as we now show. Consider the following two-level population model:

$$\begin{aligned} \text{Level 1: } u_i | t_i &\sim \varphi_p(u_i | t_i; \theta_1), \quad i = 1, \dots, M \\ \text{Level 2: } Y_{ij} | (u_i, x_{ij}) &\sim f_p(y_{ij} | x_{ij}, u_i; \theta_2), \quad j = 1, \dots, N_i, \end{aligned} \quad (2.9)$$

where φ_p and f_p denote the first and second-level *pdfs* with known covariates t_i and x_{ij} , governed by the hyperparameters θ_1 and θ_2 respectively. The model (2.7) is a special case of (2.9) by which φ_p and f_p are normal *pdfs* with $t_i = 0$ (no covariates), $\theta_1 = \sigma_u^2$ and $\theta_2 = (\beta, \sigma_\varepsilon^2)$. Suppose that the sample is drawn by the following two-stage sampling process. In the first stage a sample s_1 of $m < M$ first-level units (clusters; say, schools) is selected with probabilities $\pi_i = \Pr(i \in s_1)$ that may be correlated with the random effects u_i after conditioning on the covariates t_i . In the second stage a sub-sample s_{2i} of $n_i < N_i$ second-level units (ultimate sampling units; say, pupils) is sampled from each selected first-level unit i with probabilities $\pi_{j|i} = \Pr(j \in s_{2i} | i \in s_1)$ that may be correlated with the outcomes y_{ij} after conditioning on the covariates x_{ij} . Denote by I_i and $I_{j|i}$ the first and second-stage sampling indicators. By (2.1), the two-level sample model holding for the observed data, corresponding to the population model (2.9) is,

$$\begin{aligned} \text{Level 1:} \\ f_{s_1}(u_i | t_i; \theta_1, \gamma_1) &= \frac{\Pr(I_i = 1 | u_i, t_i; \gamma_1) \varphi_p(u_i | t_i; \theta_1)}{\Pr(I_i = 1 | t_i; \theta_1, \gamma_1)} \\ \text{Level 2:} \\ f_{s_{2i}}(y_{ij} | x_{ij}, u_i; \theta_2, \gamma_2) &= \frac{\Pr(I_{j|i} = 1 | y_{ij}, x_{ij}; \gamma_2) f_p(y_{ij} | x_{ij}, u_i; \theta_2)}{\Pr(I_{j|i} = 1 | u_i, x_{ij}; \theta_2, \gamma_2)}, \end{aligned} \quad (2.10)$$

where I assume $\Pr(I_{j|i} = 1 \mid y_{ij}, u_i, x_{ij}; \gamma_2) = \Pr(I_{j|i} = 1 \mid y_{ij}, x_{ij}; \gamma_2)$.

Remark 3. By the independence result in Remark 2, if $y_{ij} \mid u_i$ are independent under the population model, they are asymptotically independent under the sample model. Similarly, if the random effects u_i are independent under the population model, they are asymptotically independent under the sample model. Thus, the sample model (2.10) is a genuine two-level model, although with different distributions and possibly more parameters. Evidently, the models (2.9) and (2.10) are different, unless $\Pr(I_{j|i} = 1 \mid y_{ij}, x_{ij}) = \Pr(I_{j|i} = 1 \mid u_i, x_{ij})$ and $\Pr(I_i = 1 \mid u_i, t_i) = \Pr(I_i = 1 \mid t_i)$.

So far I assumed implicitly full response. Suppose, for example, that in sampled cluster (first level unit) i only a sub-sample $r_{2i} \subset s_{2i}$ respond, and denote by $R_{j|i}$ the response indicator. The second-level model for the observed outcomes is now,

Level 1:

$$\begin{aligned} f_{o2i}(y_{ij} \mid x_{ij}, u_i; \theta_2, \gamma_2, \gamma_2^*) \\ = f(y_{ij} \mid x_{ij}, u_i, I_{j|i} = 1, R_{j|i} = 1) \\ = \frac{\Pr(R_{j|i} = 1 \mid y_{ij}, x_{ij}, I_{j|i} = 1; \gamma_2^*) f_{s_{2i}}(y_{ij} \mid x_{ij}, u_i; \theta_2, \gamma_2)}{\Pr(R_{j|i} = 1 \mid x_{ij}, u_i, I_{j|i} = 1; \theta_2, \gamma_2, \gamma_2^*)}. \quad (2.11) \end{aligned}$$

The *pdf* (2.11) coupled with the level 1 *pdf* in (2.10) defines the model holding for the observed data in the case of informative cluster sampling and NMAR nonresponse.

3. How can we estimate population models from complex survey data?

In this section I review the main approaches proposed in the literature to deal with the special features of complex survey data discussed in Section 2, and propose some modifications. In order to simplify the discussion, I consider the following set up used for the simulation study in Section 4.

3.1 Population model and sampling design

Consider a stratified population $U = U_1 \cup \dots \cup U_H$ of size N . Specifically, define for every unit $j \in U$ a random vector stratification indicator $z_j = (z_{1j}, \dots, z_{Hj})'$ such that $\Pr(z_{hj} = 1) = p_h$, $\sum_{h=1}^H p_h = 1$ and $j \in U_h$ if $z_{hj} = 1$. The stratification is carried out independently between the units. Values of an outcome variable Y are generated as $y_j = \beta_0 + \beta_1 x_j + \alpha_0 \zeta_j + \alpha_1 \zeta_j x_j + \varepsilon_j$; $\varepsilon_j \sim N(0, \sigma^2)$, where the x_j 's are fixed scalar covariates, $(\beta_0, \beta_1, \alpha_0, \alpha_1)$ are fixed coefficients and

$$\zeta_j = \frac{1}{H} \sum_{h=1}^H \frac{z_{hj}}{p_h} - 1.$$

Notice that ζ_j is a random variable with mean zero and variance

$$V_\zeta = \left(\frac{1}{H^2} \sum_{h=1}^H \frac{1}{p_h} \right) - 1,$$

implying that for given covariates x_j, x_k ,

$$\begin{aligned} E_p(y_j \mid x_j) &= \beta_0 + \beta_1 x_j, \text{Var}_p(y_j \mid x_j) \\ &= (\alpha_0 + \alpha_1 x_j)^2 V_\zeta + \sigma^2, \text{Cov}_p(y_j, y_k \mid x_j, x_k) \\ &= 0, j \neq k. \end{aligned} \quad (3.1)$$

However, for unit $j \in U_h$,

$$\begin{aligned} y_j \mid x_j, z_{hj} = 1 &\sim N[(\beta_0 + \alpha_0 \zeta_h) \\ &+ (\beta_1 + \alpha_1 \zeta_h) x_j, \sigma^2]; \zeta_h = [(1/Hp_h) - 1]. \end{aligned} \quad (3.2)$$

Thus, the regression model in each stratum is the classical linear model with constant variance, but the intercepts and slopes change across the strata.

The model defined by (3.1) and (3.2) is a realistic random coefficients regression model, which I think mimics many populations encountered in practice.

We used systematic probability proportional to size (PPS) sampling within the strata for drawing the samples with the size variable defined as $z_j^* = \max\{\min[(|q_j|)^{1.5}, 9], 1\}$; $q_j \sim N(1 + x_j, 1)$. There is nothing novel about the choice of this size variable except that it allows for a clear distinction between the variance of the various estimators. This size z_j^* does not depend on the outcome y_j , and hence the sampling process within each stratum is non-informative. However for disproportionate allocation of the sample between the strata, the sampling scheme is informative because of the different models operating in different strata, such that the observed outcomes carry information on the strata membership and $\Pr(j \in s \mid y_j, x_j) \neq \Pr(j \in s \mid x_j)$. We focus on the estimation of the regression coefficients (β_0, β_1) in (3.1) as the target of inference and assume that the available sample information consists of the observed outcomes and covariates, the strata membership vectors z_{hj} and the strata sizes, $\{N_h\}$.

3.2 Including the design variables among the covariates

As implied by (2.3), the population model (*pdf*), $f_p(y_s \mid x_s)$ and the sample model $f_s(y_s \mid x_s)$ are the same if $\Pr(A_s = 1 \mid y_s, x_s) = \Pr(A_s = 1 \mid x_s) \forall y_s$. By (2.2), the response process is ignorable if $\Pr(R_i = 1 \mid y_i, x_i, I_i = 1) = \Pr(R_i = 1 \mid x_i, I_i = 1) \forall y_i$. Thus, a possible

way to account for the sampling and response effects is to add to the model covariates all the variables and interactions determining the sample and response probabilities and then integrate them out in order to estimate the model of interest. Denote these variables by $J = Z \cup L$ with population values J_U , where L defines the variables explaining the response probabilities. Assuming $f_p(y_s | x_U, j_U) = f_p(y_s | x_s, j_U)$, the use of this approach requires to fit first the model

$$f_p(y_s | x_s, J_U = j_U) = \int f_p(y_s, y_{\bar{s}} | x_U, j_U) dy_{\bar{s}}, \quad (3.3)$$

and then integrate,

$$f_p(y_s | x_s) = \int f_p(y_s | x_s, j_U) f_p(j_U | x_s) dj_U. \quad (3.4)$$

Variants of this approach can be found in DeMets and Halperin (1977), Holt, Smith and Winter (1980), Nathan and Holt (1980), Jewell (1985), Skinner (1994), Chambers and Skinner (2003, Chapter 2) and Gelman (2007).

The use of the approach is appealing, and it has the advantage of allowing classical model based inference procedures once the variables $J_U = Z_U \cup L_U$ are included in the model, but it is often limited in practice for the following reasons:

1. It requires knowledge of the population values of all the variables determining the sample selection and response, and this information is usually unknown to the analyst fitting the model because of confidentiality restrictions or other reasons. Even if known, including in the model all the geographic and operational variables used for the sampling design and the variables explaining the response may be formidable.
2. In practice there may be many covariates and many design variables, and modelling the relationship between the design variables and the covariates in order to integrate out the effect of the former variables can be complicated and may no longer reproduce the original target model.

Feder (2011) proposes the following simple solution to this problem. Suppose first that the design variables and the covariates are known for every element in the population. The proposed solution consists of imputing the missing population outcomes using the model $f_p(y_s | x_s, J_U = j_U)$ fitted to the sample data, and then fitting the population model $f_p(y_j | x_j)$ using all the population values, with the missing outcomes replaced by their imputed values. When the design variables and the covariates are unknown for the non-sampled units, they need to be imputed as well. The imputation may be carried out by sampling with replacement $(N - n)$ values (x_i, z_i) from the sample values with probabilities $\tilde{p}_i = (w_i - 1) / \sum_{k=1}^n (w_k - 1)$ on each draw, where the w_i 's are the sampling weights. See Pfeffermann

and Sikov (2011) for justification of this procedure under the sample model and an extension for the case of NMAR nonresponse.

3. The approach is not operational when the inclusion in the sample depends also on the outcome values, that is, $Z_U = \{Y_U, Z_U^*\}$ and $\Pr(A_s = 1 | Y_U, X_U, Z_U^*) \neq \Pr(A_s = 1 | X_U, Z_U^*)$. A classical example is *case-control studies* (Scott and Wild 2009), but a similar problem arises when the nonresponse is NMAR.

Remark 4. Including the design variables and the variables explaining the response in the model does not necessarily require integrating them out even if they are not part of the covariates of interest, as the following example shows.

Example 4: Suppose that a sample of size n is selected with probabilities defined by the population values of design variables Z and that all the sampled units respond. Let the population distribution of Y, X, Z be multivariate normal. The data available to the analyst consist of the sample values $[y_s, x_s]$ and the population values Z_U . Using properties of the multivariate normal distribution, $E_p(y | x) = \beta_0 + \beta_{yx}x$ for some coefficients (β_0, β_{yx}) , but the OLS estimate of β_{yx} is biased because the sampling probabilities depend on Z , which is correlated with Y . The *mle* of β_{yx} for the case of a trivariate normal distribution is (DeMets and Halperin 1977),

$$\hat{\beta}_{yx} = \left\{ s_{xy} + \frac{s_{yz}s_{xz}}{s_{zz}} \left(\frac{\hat{\sigma}_z^2}{s_{zz}} - 1 \right) \right\} / \left\{ s_{xx} + \frac{s_{xz}^2}{s_{zz}} \left(\frac{\hat{\sigma}_z^2}{s_{zz}} - 1 \right) \right\}, \quad (3.5)$$

where $s_{uv} = n^{-1} \sum_{i=1}^n (u_i - \bar{u}_s)(v_i - \bar{v}_s)$ and $\hat{\sigma}_z^2 = N^{-1} \sum_{i=1}^N (z_i - \bar{z}_U)^2$, with \bar{u}_s, \bar{v}_s and \bar{z}_U defining the corresponding sample and population means. Thus, the population values of Z feature in this case in the optimal estimator of the target parameter β_{yx} . Holt *et al.* (1980) extend this result to the case where Y, X, Z are vector variables. Nathan and Holt (1980) establish conditions under which $\hat{\beta}_{yx}$ is consistent without the multivariate normality assumptions. Pfeffermann and Holmes (1985) study the robustness of the estimator to model misspecification.

3.3 Using the sampling weights as surrogate for the design variables

For situations where there are too many design variables determining the sample selection to include them all in the model, or when some or all of these variables are unknown to the analyst, it is often advocated to include in the model the sampling weights as surrogate of the design variables. Examples of the use of this approach can be found in DuMouchel and Duncan (1983), Särndal and Wright

(1984), Rubin (1985), Chambers, Dorfman and Wang (1998) and Wu and Fuller (2006).

Rubin (1985) defines the vector $a = (a_1, \dots, a_N)' = a(Z_U)$ to be an adequate summary of Z_U if $\Pr(A_s = 1 | Z_U) = \Pr(A_s = 1 | a)$. The author shows that the vector $\pi_U = (\pi_1, \dots, \pi_N)$ of the sample inclusion probabilities is the coarsest possible adequate summary of Z_U , though it may be too coarse. It follows therefore that for sampling designs such that $\Pr(A_s = 1 | Y_U, Z_U) = \Pr(A_s = 1 | Z_U)$, if π_U is an adequate summary, the sample selection can be ignored for inference on the parameters of $f_p(y_s | x_s, \pi_U)$. In order to estimate the target model $f_p(y | x)$ in this case, one can follow the same steps as in Section (3.2) with π_U taking the role of Z_U .

The use of this approach reduces the dimension of the added covariates but it requires knowledge of the sample inclusion probabilities (or the sampling weights) for all the population units, which may not be available in the case of a secondary analysis. The case of nonresponse is particularly problematic since the response probabilities are generally unknown and need to be estimated. Another major problem with this approach is that for general sampling designs, the vector π_U may not be an adequate summary of Z . Sugden and Smith (1984) and Smith (1988) establish necessary design information required for sampling ignorability.

Remark 5. Even though the vector π_U is not always an adequate summary of Z_U , for sampling designs such that $\Pr(I_i = 1 | y_i, x_i, \pi_i) = \pi_i$, $f_s(y_i | x_i, \pi_i) = f_p(y_i | x_i, \pi_i)$, so that the marginal population and sample *pdfs* for a given sampled unit are nonetheless the same when adding π_i to the covariates (see Skinner 1994).

Remark 6. In the empirical set up described in Section 3.1 there is a one to one correspondence between the design variables (z'_j, z_j^*) and the sampling weights (w_h, w_j) .

3.4 Methods based on probability weighting

So far we considered methods requiring knowledge of the variables J determining the sample selection and response probabilities, or at least an adequate summary of them. The methods considered below only require knowledge of the sampling weights for the responding sampled units. As such, they are restricted to situations of full response, or to cases where the response probabilities can be estimated sufficiently accurately, in which case the sampling weight for a responding unit is the inverse of the product of the unit's selection probability and its estimated response probability. Probability weighting (PW) is discussed in numerous articles; see the recent discussion in Pfeffermann and Sverchkov (2009) and the references therein. As before, we focus here on estimation of population models.

To introduce the idea, consider the case of a *census* with full response. Assuming independent outcomes, the model parameters, θ , are typically estimated in this case by solving *census* estimating equations of the form,

$$\sum_{j=1}^N u(y_j, x_j; \theta) = 0. \quad (3.6)$$

In the case of *mle*, $u(y_j, x_j; \theta) = (\partial / \partial \theta) \log f_p(y_j | x_j; \theta)$, the j^{th} score. In practice, data are available for only a sample $s \subset U$ and the equations (3.6) are replaced by their randomization unbiased Horvitz-Thompson estimator,

$$\sum_{i \in s} w_i u(y_i, x_i; \theta) = 0, \quad (3.7)$$

where the w_i 's are the sampling weights.

Remark 7. When the census estimating equations (3.6) are the likelihood equations, the estimators obtained by solving (3.7) are known in the sampling literature as 'pseudo *mle*' (*pmle*). See Binder (1983), Skinner *et al.* (1989), Pfeffermann (1993, 1996) and Godambe and Thompson (2009) for discussion with many examples. This approach is implemented in many software packages such as SAS, STATA, SUDAAN, *etc.*

Example 5. In the case of the standard linear regression model, the *pmle* or PW estimator of the vector coefficient β solves the equations $\sum_{i \in s} w_i (y_i - x'_i \hat{\beta}_{pw}) x_i = 0$;

$$\hat{\beta}_{pw} = \left[\sum_{i \in s} w_i x_i x'_i \right]^{-1} \sum_{i \in s} w_i x_i y_i. \quad (3.8)$$

The PW estimator of the residual variance is $\hat{\sigma}_{pw}^2 = \sum_{i \in s} w_i (y_i - x'_i \hat{\beta}_{pw})^2 / (\sum_{i \in s} w_i - k)$, where $k = \dim(\beta)$.

For logistic regression, the pseudo likelihood equations (with no explicit solution) are,

$$\begin{aligned} \sum_{i \in s} w_i [y_i - \tilde{p}_i(x_i)] x_i &= 0; \quad \tilde{p}_i(x_i) \\ &= \Pr_p(y_i = 1 | x_i) \\ &= \exp(x'_i \beta) / [1 + \exp(x'_i \beta)]. \end{aligned} \quad (3.9)$$

Example 6. Let $u(y_j; \theta) = [\Delta(\theta - y_j) - F_p(\theta)]$ where $F_p(\theta)$ is the cumulative population distribution at θ and $\Delta(a) = 1(0)$ when $a \geq 0$ ($a < 0$). The PW estimator of $F_p(\theta)$ is $\hat{F}_{p, pw}(\theta) = \sum_{i \in s} w_i \Delta(\theta - y_i) / \sum_{i \in s} w_i$, the familiar Hájek (1971) estimator.

The notable property of PW estimators is that they are generally $r - p$ consistent. (See Section 2.2 for definition of the $r - p$ distribution). This can be seen by decomposing $(\hat{\theta}_{pw} - \theta) = (\hat{\theta}_{pw} - \hat{\theta}_{cen}) + (\hat{\theta}_{cen} - \theta)$, where $\hat{\theta}_{cen}$ is the (hypothetical) solution of the census equations (3.6). Under general conditions, $(\hat{\theta}_{pw} - \hat{\theta}_{cen}) = O_p(n^{-0.5})$ and $(\hat{\theta}_{cen} - \theta) = O_p(N^{-0.5})$, thus establishing the $r - p$ consistency of $\hat{\theta}_{cen}$ under these conditions. The $r - p$ variance of $\hat{\theta}_{pw}$ can be decomposed as,

$$\text{Var}_{r-p}(\hat{\theta}_{pw}) = E_p[\text{Var}_r(\hat{\theta}_{pw})] + \text{Var}_p[E_r(\hat{\theta}_{pw})]. \quad (3.10)$$

For single stage sampling, if n is much smaller than N as is usually the case, the second term on the right hand side of (3.10) is negligible compared to the first term, and $\text{Var}_{r-p}(\hat{\theta}_{pw})$ can be estimated by the randomization variance estimator $\hat{\text{Var}}_r(\hat{\theta}_{pw})$. This result does not necessarily hold for cluster sampling since in this case $\text{Var}_r(\hat{\theta}_{pw})$ is typically of order $O(1/m)$ where m is the number of sampled clusters, and under a suitable model $\text{Var}_p[E_r(\hat{\theta}_{pw})]$ is $O(1/M)$ where M is the number of population clusters. For $\hat{\text{Var}}_r(\hat{\theta}_{pw})$ to be an adequate estimator of $\text{Var}_{r-p}(\hat{\theta}_{pw})$ in this case, m must be much smaller than M .

Remark 8. The consistency of PW estimators under correct population model specification may also be established under the sample distribution (Equation 2.1). Consider the estimator $\hat{\beta}_{pw}$ in (3.8) and write $\hat{\beta}_{pw} = \beta + [\sum_{i \in s} w_i x_i x_i']^{-1} \sum_{i \in s} w_i x_i \varepsilon_i$ where the ε_i 's are the population model residuals. The key result leading to the consistency of $\hat{\beta}_{pw}$ under the sample distribution is that if $\Pr(I_i = 1 | y_i, x_i, \pi_i) = \pi_i$ then $E_s(w_i \varepsilon_i) = E_s(w_i)E_p(\varepsilon_i) = 0$ (follows from 3.14 below). In fact, by viewing the covariates as random with (y_i, x_i) having some joint distribution,

$$\beta = \arg \min_{\beta} E_p(y_i - x_i' \beta)^2 = \arg \min_{\beta} E_s[w_i (y_i - x_i' \beta)^2],$$

implying that $\hat{\beta}_{pw}$ is the optimal estimator (in weighted least-squares metric) of β under the sample distribution of (y_i, x_i) . See also (3.24) below. Godambe and Thompson (1986, 2009) establish and discuss other optimality properties of estimators solving estimating equations of the form $\sum_{i \in s} w_i u(y_i, x_i; \theta) = 0$. The following example shows how probability weighting can be used when modelling clustered populations.

Example 7. Consider the population two-level (random intercept) model,

Level 1:

$$u_i \sim N(t_i' \gamma, \sigma_u^2), \quad i = 1, \dots, M \quad (3.11)$$

Level 2:

$$y_{ij} = x_{ij}' \beta + u_i + \varepsilon_{ij}, \quad \varepsilon_{ij} \sim N(0, \sigma_\varepsilon^2), \quad j = 1 \dots N_i$$

where ε_{ij} and u_i are independent for all i and j . The unknown parameters are the vectors of coefficients $\vartheta = (\beta', \gamma')'$ and the variances $\tau = (\sigma_\varepsilon^2, \sigma_u^2)'$. Assume full response. Under ignorable sampling of first and second-level units, the *mle* of (ϑ, τ) is computed conveniently by iterating between the estimation of ϑ for 'known' τ and the estimation of τ for 'known' ϑ , with the 'known' values defined by the estimators from the previous iteration. The two sets of estimators on the r^{th} iteration are the solutions of linear equations of the form, $P^{(r)} \vartheta = q^{(r)}$, $R^{(r)} \tau = s^{(r)}$,

with appropriate definition of the matrices $(P^{(r)}, R^{(r)})$ and the vectors $(q^{(r)}, s^{(r)})$, $r = 1, 2, \dots$, (Goldstein 1986). When applied to all the population values, these equations define the *census* estimating equations.

Suppose, as before, that a sample s_1 of first-level units is sampled with probabilities $\pi_i = \Pr(i \in s_1)$, and that subsamples s_{2i} of size $n_i < N_i$ are sampled from each selected first-level unit i with probabilities $\pi_{ji} = \Pr(j \in s_{2i} | i \in s_1)$. The *pml* for this model can be obtained by first expressing the elements of the matrices $(P^{(r)}, R^{(r)})$ and the vectors $(q^{(r)}, s^{(r)})$ as sums over first and second-level units, and then estimating each population sum of the form $\sum_{i=1}^M d_i$ by the H-T estimator $\sum_{i \in s_1} (d_i / \pi_i)$, and each population sum of the form $\sum_{j=1}^{N_i} d_{ij}$ by the H-T estimator $\sum_{j \in s_{2i}} (d_{ij} / \pi_{ji})$. See Pfeffermann, Skinner, Holmes, Goldstein and Rasbash (1998b). Pfeffermann and Sverchkov (2009) review other methods of probability weighting in two-level models.

Probability weighting is in broad use both for estimation of finite-population quantities, referred to in the literature as descriptive inference, and for 'analytic inference' on population models. The main attraction of this method is its simplicity. It is generally viewed as being 'model free', except when having to estimate the response probabilities, which is often based on models, and hence more robust than other methods, but when used for analytical inference, this view is questionable.

Probability-weighted estimators are randomization consistent for the corresponding descriptive population quantities (CDPQ), defined as the (hypothetical) solutions of the census estimating equations. However, if the population model is misspecified, the target CDPQ are not (model) p -consistent for the true model parameters and the PW estimators are not $r-p$ consistent either. So, probability weighting provides no protection against model misspecification, although the estimated CDPQ may be useful for various kinds of inference. See Pfeffermann (1993) and Binder and Roberts (2009) for discussion and examples.

Estimating the randomization variance of probability-weighted estimators is generally simple, utilizing available techniques in finite population sampling. Binder (1983) developed a general approach for estimating the randomization variance of estimators obtained as the solution of probability-weighted estimating equations; see also Binder and Roberts (2009) and Godambe and Thompson (2009). Fuller (1975), Binder (1983), Chambless and Boyle (1985) and Francisco and Fuller (1991) developed central limit theorems applicable to probability-weighted estimators.

In spite of these desirable properties of probability-weighting, the method has some severe limitations:

1. It is restricted mostly to point estimation. Probabilistic inference like confidence intervals or

hypothesis testing generally requires large sample normality assumptions. In particular, the randomization distribution does not lend itself to the use of classical inference methods such as likelihood-based or Bayesian inference.

2. The variances of probability-weighted estimators are computed with respect to the randomization distribution and the use of this approach does not permit conditioning on the selected sample, for example, conditioning on the observed covariates or the selected clusters in a multi-level model.
3. As often illustrated in the literature, probability-weighted estimators generally have larger variances than model-based estimators, notably for small samples and large variation of the sampling weights.
4. The use of the randomization distribution does not lend itself to prediction problems such as the prediction of the outcome for non-sampled units with known covariates under a regression model, or the prediction of small area means for areas with no samples in a small-area estimation problem.

3.5 Modifications of the sampling weights

When estimating finite population quantities, the sampling weights are often modified by imposing calibration equations, which match the PW estimators of covariates for which the population totals are known with the actual totals. The use of calibration is particularly useful in the case of nonresponse; see Kott (2009) for recent discussion with references. We later discuss the use of *empirical likelihood* for analytical inference on population models, which also attempts to incorporate calibration equations, although in a different manner. Below, I review two modifications of the sampling weights aimed at reducing the variances of the weighted estimators of model parameters under the *sample distribution* (2.1). A combination of the two modifications is also considered.

Magee (1998) considers a linear regression model but the results can be extended to other population models. The author shows that under certain moment assumptions, any estimator $\hat{\beta}_{\text{mg}}(a) = [\sum_{i \in s} w_i a_i(\alpha) x_i x_i']^{-1} \sum_{i \in s} w_i a_i(\alpha) x_i y_i$ with positive weights $a_i(\alpha) = a(x_i, \alpha)$ is consistent for β under the sample distribution. The weights $a(x_i, \alpha)$ belong to a parameterized family of functions with the vector parameter α chosen to minimize a scalar variance criterion such as the determinant or the trace of the asymptotic variance estimator,

$$\begin{aligned} & A \hat{\text{var}}[\hat{\beta}_{\text{mg}}(a)] \\ &= \left[\sum_{i \in s} w_i a_i(\alpha) x_i x_i' \right]^{-1} \sum_{i \in s} w_i^2 a_i^2(\alpha) \hat{\varepsilon}_i^2 x_i x_i' \\ & \quad \left[\sum_{i \in s} w_i a_i(\alpha) x_i x_i' \right]^{-1}, \end{aligned} \quad (3.12)$$

where $\hat{\varepsilon}_i = (y_i - x_i' \hat{\beta}_{\text{pw}})$. The choice of the function $a(x_i, \alpha)$ is up to the analyst but the obvious idea is to choose a function that is believed to be approximately inversely proportional to the residual variance under the sample model. The resulting 'Quasi-Aitken' estimator is shown to have asymptotically a lower variance under the sample distribution than the probability-weighted estimator $\hat{\beta}_{\text{pw}}$. Recall from Remark 8 that $\hat{\beta}_{\text{pw}}$ is consistent for β under the sample distribution, justifying comparing the asymptotic variances of the two estimators under this distribution.

Pfeffermann and Sverchkov (1999) propose another modification. Consider the population model,

$$y_j = m(x_j; \theta) + \varepsilon_j, \quad E_p(\varepsilon_j | x_j) = 0, \quad E_p(\varepsilon_j^2 | x_j) = \sigma^2, \quad (3.13)$$

where $m(x_j; \theta)$ has a known form. Let $q_i = w_i / E_s(w_i | x_i)$. The authors show that if $\Pr(I_i = 1 | \pi_i, y_i, x_i) = \pi_i$,

$$E_p(y_i | x_i) = E_s(w_i y_i | x_i) / E_s(w_i | x_i). \quad (3.14)$$

Thus, for vectors $\tilde{\theta}$ in the plausible parameter space Θ ,

$$\begin{aligned} \theta &= \underset{\tilde{\theta}}{\text{argmin}} \frac{1}{n} \sum_{i \in s} E_p \{ [y_i - m(x_i; \tilde{\theta})]^2 | x_i \} \\ &= \underset{\tilde{\theta}}{\text{argmin}} \frac{1}{n} \sum_{i \in s} E_s \{ q_i [y_i - m(x_i; \tilde{\theta})]^2 | x_i \}. \end{aligned}$$

The vector θ can be estimated therefore by solving the minimization problem,

$$\begin{aligned} \hat{\theta}_q &= \underset{\tilde{\theta}}{\text{argmin}} \frac{1}{n} \sum_{i=1}^n \hat{q}_i [y_i - m(x_i; \tilde{\theta})]^2; \\ \hat{q}_i &= w_i / \hat{E}_s(w_i | x_i). \end{aligned} \quad (3.15)$$

The use of this estimator requires estimating $E_s(w_i | x_i)$ but under mild regularity conditions $\hat{\theta}_q$ is consistent for θ even when the expectation $E_s(w_i | x_i)$ is misspecified. See Pfeffermann and Sverchkov (2009) and Section 4.1 of this paper for examples of the specification and estimation of $E_s(w_i | x_i)$.

Example 8. Under the linear regression population model with constant variance,

$$\hat{\beta}_q = \left[\sum_{i \in s} \hat{q}_i x_i x_i' \right]^{-1} \sum_{i \in s} \hat{q}_i x_i y_i. \quad (3.16)$$

As easily verified, $\hat{\beta}_q$ is randomization consistent for the census regression coefficients $\tilde{B} = [\sum_{j=1}^N x_j x_j' / E_s(w_j | x_j)]^{-1} \sum_{j=1}^N x_j y_j / E_s(w_j | x_j)$, and hence $p-r$ consistent for β , even when $E_s(w_i | x_i)$ is misspecified.

The obvious difference between the PW estimator $\hat{\theta}_{pw}$ and the estimator $\hat{\theta}_q$ is that the latter estimator uses the adjusted weights $q_i = w_i / \hat{E}_s(w_i | x_i)$. When the sample selection depends only on the covariates, the sampling process is ignorable. Hence, to protect against informative sampling, it is only necessary to account for the net sampling effects on the target conditional *pdf* of $y_i | x_i$. This is achieved by using the weights q_i . In contrast, the sampling weights w_i account for the sampling effects on the joint distribution of (y_i, x_i) . As a result, they tend to be more variable and the estimator $\hat{\theta}_{pw}$ has a larger variance.

A combination of the last two modifications is also possible and examined in Section 4. The simple idea proposed by Dr. Moshe Feder (private communication) is to apply the modification of Magee (1998) to the estimator $\hat{\beta}_q$ instead of the estimator $\hat{\beta}_{pw}$, that is, use the estimator,

$$\hat{\beta}_{mg-q}(a) = \left[\sum_{i \in s} \hat{q}_i a_{i,q}(\alpha) x_i x_i' \right]^{-1} \sum_{i \in s} \hat{q}_i a_{i,q}(\alpha) x_i y_i, \quad (3.17)$$

where the vector parameter α is now chosen to minimize a scalar variance criterion of the asymptotic variance estimator, $\text{Avar}[\hat{\beta}_{mg-q}(a)]$, computed similarly to (3.12).

3.6 Likelihood based methods

3.6.1 Use of the sample model for maximum likelihood estimation

A natural way of estimating the population model parameters is by maximization of the sample likelihood. Assume first full response and that the sample observations are independent under the sample distribution. The likelihood has then the form,

$$L_s(\theta, \gamma; y_s, x_s) = \prod_{i \in s} \frac{\Pr(I_i = 1 | x_i, y_i; \gamma) f_p(y_i | x_i; \theta)}{\Pr(I_i = 1 | x_i; \gamma, \theta)}. \quad (3.18)$$

As before, we assume $\Pr(I_i = 1 | \pi_i, y_i, x_i) = \pi_i$, implying $\Pr(I_i = 1 | x_i, y_i) = E_p(\pi_i | x_i, y_i)$. By (3.14), The sample likelihood can be written therefore as,

$$L_s(\theta, \gamma; y_s, x_s) = \prod_{i \in s} \frac{E_s(w_i | x_i; \gamma, \theta) f_p(y_i | x_i; \theta)}{E_s(w_i | y_i, x_i; \gamma)}. \quad (3.19)$$

The expectations on the right hand side of (3.19) are with respect to the sample *pdf* of the sampling weights. Thus,

when the weights are known for the sampled units as is usually the case under full response, the expectations can be modelled and estimated by regressing w_i against (y_i, x_i) , using classical model fitting procedures. Suppose first that the weights are continuous such as in probability proportional to size (PPS) sampling with a continuous size variable. For a given form of the population model, the expectations $E_s(w_i | y_i, x_i; \gamma)$ and $E_s(w_i | x_i; \gamma, \theta)$ can be obtained then in two steps:

1. Identify and estimate $\hat{E}_s(w_i | y_i, x_i; \gamma) = E_s(w_i | y_i, x_i; \hat{\gamma})$, using the sample data.
2. Integrate $\int [1/E_s(w_i | y, x_i; \hat{\gamma})] f_p(y | x_i; \theta) dy$ to obtain $E_p(\pi_i | x_i; \theta; \hat{\gamma})$. Compute, $\hat{E}_s(w_i | x_i; \theta, \hat{\gamma}) = 1/E_p(\pi_i | x_i; \theta, \hat{\gamma})$ (follows from 3.14).

Estimating the vector parameter γ outside the likelihood and then substituting the estimate in (3.19) and maximizing the likelihood as a function of the vector parameter θ only, usually yields more stable results than maximizing the likelihood over (θ, γ) simultaneously.

Estimation of the expectations $E_s(w_i | y_i, x_i; \gamma)$ and $E_s(w_i | x_i; \theta, \gamma)$ in the case of discrete inclusion probabilities is similar.

Example 9. Consider the case of multinomial-logistic regression with a discrete covariate x and M possible values of the outcome y . Assuming that $E_s(w_i | y_i = m, x_i = k)$ is not a function of the model parameters, it can be estimated by \bar{w}_{mk} , the mean of the weights in cell (m, k) , and thence $\hat{\pi}_{mk} = \hat{\Pr}_p(i \in s | y_i = m, x_i = k) = (1 / \bar{w}_{mk})$. We obtain:

$$\Pr_s(y_i = m | x_i = k; \theta) \cong \frac{[\Pr_p(y_i = m | x_i = k; \theta) / \bar{w}_{mk}]}{\sum_{m^*=1}^M [\Pr_p(y_i = m^* | x_i = k; \theta) / \bar{w}_{m^*k}]} \quad (3.20)$$

The sampling weights feature in the sample model, but this is not an application of classical probability weighting. Notice that with this approximation the parameters in the population and the sample model are the same. In our empirical study we use a similar approximation for the sample distribution by categorizing the values of a continuous outcome. See Pfeffermann and Sverchkov (1999) for other examples.

Next consider the estimation of the vector parameter θ governing the population model. Under mild conditions, θ is the unique solution of the equations,

$$W_U(\theta) = \sum_{j \in U} E_p(\delta_j | x_j) = 0; \quad \delta_j = (\delta_{j,0}, \delta_{j,1}, \dots, \delta_{j,k})' = \partial \log f_p(y_j | x_j; \theta) / \partial \theta. \quad (3.21)$$

Pfeffermann and Sverchkov (2003) consider three different approaches for estimating θ . The common feature of these

approaches is that the only data used for estimation are the observations $\{(y_i, x_i, w_i), i \in s\}$, similarly to the PW estimators and their modifications considered in Section 3.5. In Section 3.6.2 we consider the use of the ‘full likelihood’, which assumes knowledge of the covariates $\{x_j, j \in U\}$, and possibly also additional design information.

The first approach redefines the parameter equations with respect to the sample model. Assuming that $E_s(w_i | x_i; \theta, \gamma)$ in (3.19) is differentiable with respect to θ , the sample model parameter equations are $W_{1s}(\theta) = \sum_{i \in s} E_s \{ [\partial \log f_s(y_i | x_i; \theta, \gamma) / \partial \theta] | x_i \} = \sum_{i \in s} E_s \{ [\delta_i + \partial \log E_s(w_i | x_i; \theta, \gamma) / \partial \theta] | x_i \} = 0$. The vector θ is estimated under this approach by solving the equations,

$$W_{1s,e}(\theta) = \sum_{i \in s} [\delta_i + \partial \log E_s(w_i | x_i; \theta, \gamma) / \partial \theta] = 0. \quad (3.22)$$

The second approach applies the relationship (3.14) to the parameter equations (3.21). For a random sample from the sample model, the equations are now $W_{2s}(\theta) = \sum_{i \in s} E_s(q_i \delta_i | x_i) = 0$, where $q_i = w_i / E_s(w_i | x_i)$. The vector θ is estimated under this approach by solving the equations,

$$W_{2s,e}(\theta) = \sum_{i \in s} q_i \delta_i = 0. \quad (3.23)$$

The third approach uses the property that if θ solves (3.21), then it solves also the equations, $\tilde{W}_U(\theta) = \sum_{j \in U} E_p(\delta_j) = E_x[\sum_{j \in U} E_p(\delta_j | x_j)] = 0$, where $E_x(\cdot)$ is the expectation of x (which is viewed as random) with respect to the population distribution. Hence, by (3.14), for a random sample from the sample model, the parameter equations are $W_{3s}(\theta) = \sum_{i \in s} E_s(w_i \delta_i) = 0$, with estimating equations,

$$W_{3s,e}(\theta) = \sum_{i \in s} w_i \delta_i = 0. \quad (3.24)$$

Note that the equations (3.24) are the *pseudo-likelihood* equations (Remark 7).

Remark 9. The use of the weights $q_i = w_i / E_s(w_i | x_i)$ for population model parameter estimation has been justified already in Section 3.5 by reference to least-squares estimation. See the discussion in that section regarding the difference between the use of the weights q_i and the weights w_i . Pfeffermann and Sverchkov (1999, 2003) illustrate that estimating θ by solving the equations (3.23) yields estimators with lower randomization variance than estimating θ by solving the equations (3.24). Notice that under the assumption of a linear regression model operating in the population, the solution of (3.24) yields the PW estimator (3.8), and the solution of (3.23) yields the q -weighted estimator (3.16).

Remark 10. The use of the sample model for estimation of multi-level population models is considered in Pfeffermann, Moura and Nascimento-Silva (2006), using the Bayesian

approach. Pfeffermann and Sverchkov (2007) fit multi-level models for small area estimation under informative sampling of areas and within the areas, following the frequentist approach.

So far we assumed full response. Next consider the case of NMAR nonresponse. In this case the response process needs to be modelled as well. By (2.2) and with added parameter notation the ‘respondents’ likelihood takes the form,

$$L_{\theta} = \prod_{i=1}^r f(y_i | x_i, I_i = 1, R_i = 1; \theta^*, \gamma^*) \\ = \prod_{i=1}^r \frac{\Pr(R_i = 1 | y_i, x_i, I_i = 1; \gamma^*) f_s(y_i | x_i; \theta^*)}{\Pr(R_i = 1 | x_i, I_i = 1; \gamma^*, \theta^*)}, \quad (3.25)$$

where $\theta^* = (\theta, \gamma)$ represents the parameters of the sample distribution under full response (Equation 3.19), and γ^* represents the parameters of the response process. Notice that unlike the sampling probabilities $\pi_i = \Pr(i \in s)$, which are generally known and can be used for estimating the probabilities $\Pr(I_i = 1 | y_i, x_i; \gamma)$ as explained before, the response probabilities are generally unknown.

Chang and Kott (2008) propose a method of estimating the response probabilities, which uses known totals of calibration variables. The authors assume a parametric model for the response probabilities that may depend on the outcome value, and estimate the unknown parameters of this model by regressing the totals of the calibration variables against their H-T estimators. The weights used for the H-T estimators are the product of the sampling weights and the inverse of the response probabilities under the model. Let c_i define the values of the calibration variables for unit i and denote $p(y_i, x_i; \gamma^*) = \Pr(R_i = 1 | y_i, x_i, I_i = 1; \gamma^*)$. Chang and Kott (2008) estimate the unknown parameters by setting the nonlinear regression equations,

$$C^U = \sum_{i=1}^r w_i \frac{c_i}{p(y_i, x_i; \gamma^*)} + \varepsilon^*,$$

where $C^U = \sum_{j=1}^N c_j$ and ε^* is a vector of errors. The parameters γ^* are estimated by the iterative algorithm

$$\hat{\gamma}^{(j+1)} = \hat{\gamma}^{(j)} + \left\{ \hat{H}(\hat{\gamma}^{(j)})^T V^{-1}(\hat{\gamma}^{(j)}) \hat{H}(\hat{\gamma}^{(j)}) \right\}^{-1} \\ \hat{H}(\hat{\gamma}^{(j)})^T V^{-1}(\hat{\gamma}^{(j)}) \left(C^U - \sum_{i=1}^r w_i \frac{c_i}{\pi(y_i, v_i; \hat{\gamma}^{(j)})} \right), \quad (3.26)$$

where

$$\hat{H}(\hat{\gamma}^{(j)}) = \frac{\partial \left[\sum_{i=1}^r w_i \frac{c_i}{\pi(y_i, v_i; \gamma)} \right]}{\partial \gamma} \bigg|_{\gamma = \hat{\gamma}^{(j)}} \text{ and } V^{-1}(\hat{\gamma}^{(j)})$$

is the inverse of the estimated quasi-randomization variance of

$$\sum_{i=1}^r w_i \frac{c_i}{\pi(y_i, v_i; \gamma)},$$

computed at $\gamma = \hat{\gamma}^{(j)}$.

Chang and Kott (2008) do not assume a model for the outcome and their approach is therefore restricted to estimation of the model for the response probabilities. Pfeffermann and Sikov (2011) use the likelihood (3.25) for estimating population models assuming noninformative sampling. Maximization of the likelihood is carried out by iterating between maximization of the likelihood with respect to θ^* for given γ^* , and the solution of calibration equations with respect to γ^* for given θ^* , using known totals of calibration variables, similarly to Chang and Kott (2008). The 'given' parameters are the estimates from the previous iteration. The authors show how to estimate the distribution of the missing covariates and outcome for a nonresponding unit and use this distribution for imputing the missing outcomes and hence predicting the finite population total of the outcome variable.

Estimation of the population model by fitting the sample model has some important advantages not shared by the other approaches considered in this article.

1. Once the sample model is specified, it lends itself to standard model based inference such as likelihood based methods, Bayesian inference or semi-parametric modelling. It is important to emphasize in this regard that the goodness of fit of the postulated population model can be evaluated by testing the goodness of fit of the sample model fitted to the observed outcomes, using classical model diagnostic techniques. See Krieger and Pfeffermann (1997) and Pfeffermann and Sikov (2011) for appropriate test statistics with illustrations.
2. The sample likelihood provides a coherent way of handling NMAR nonresponse when estimating population models. Methods based on probability weighting require knowledge or good estimators of the response probabilities. The use of the full likelihood (see below) requires knowledge of the covariates of nonsampled units.
3. Application of this approach permits the use of conditional inference, given the sample of responding units, for example, conditioning on the observed covariates.
4. The models holding for the observed outcomes and the response probabilities define the model holding for the missing outcomes of the non-sampled units or the nonrespondents, which can be used for

imputation of these outcomes. Methods based on probability weighting and variants thereof allow estimating the population model but under informative sampling and NMAR nonresponse, the population model cannot be used for prediction or imputation of the missing outcomes. See Sverchkov and Pfeffermann (2004) and Pfeffermann and Sikov (2011) for illustrations.

5. The use of the sample model enables testing whether the sampling process can be ignored. Pfeffermann and Sverchkov (2009) review several test statistics proposed in the literature for testing the ignorability of the sample selection.

3.6.2 The full likelihood

Theoretically, a more efficient way of estimating the unknown population model parameters is to base the likelihood on the joint distribution of the sample data and the sample membership indicators. Under full response, the *full likelihood* is then,

$$L_I(\theta, \gamma; I_U, y_s, x_s, x_{\bar{s}}) = \prod_{i \in s} \Pr(I_i = 1 | y_i, x_i; \gamma) f_p(y_i | x_i; \theta) \prod_{j \notin s} [1 - \Pr(I_j = 1 | x_j; \theta, \gamma)], \quad (3.27)$$

where $I_U = \{I_1, \dots, I_N\}$ is the vector of sample inclusion indicators and $\Pr(I_j = 1 | x_j; \theta, \gamma) = \int \Pr(I_j = 1 | y_j, x_j, \gamma) f_p(y_j | x_j, \theta) dy_j$ is the *propensity score* of unit j . The likelihood (3.27) assumes $\Pr(I_U | y_U, x_U) = \prod_{k \in U} \Pr(I_k | y_k, x_k)$ (Poisson sampling), but it can be generalized to other sampling designs. The full likelihood has the advantage of accounting for the sampling probabilities of units outside the sample, thus utilizing more information, but it requires knowledge of the covariates of all the population units. See, for example, Gelman, Carlin, Stern and Rubin (2003) and Little (2004). Modelling the joint distribution of the covariates for units outside the sample and integrating them out of the likelihood can be very complicated in practice and is formidable when there are many of them. Pfeffermann *et al.* (2006) compare empirically the use of the sample likelihood with the use of the full likelihood for multi-level models in a Bayesian context. The two approaches yield similar results, but this of course may not be the case in other applications.

Another way of defining the full likelihood is by application of the *Missing Information Principle* (MIP, Orchard and Woodbury 1972). The basic idea is to express the sample score function as the conditional expectation of the population score function, given the sample data. Following Chambers and Skinner (2003, Chapter 2), define the *full-sample likelihood* as $L_{fs}(\lambda) = f(\lambda; y_s, x_s, I_U, z_U)$

where, as before, z_U is a known matrix of population values underlying the sample selection and λ defines the unknown model parameters. The corresponding *full-population* likelihood is $L_{fU}(\lambda) = f(\lambda; y_U, x_U, I_U, z_U)$ where $y_U = (y_s, y_{\bar{s}})$ and $x_U = (x_s, x_{\bar{s}})$. The MIP principle states that,

$$sc_s(\lambda) = (\partial / \partial \lambda) \log[L_{fs}(\lambda)] \\ = E_p[(\partial / \partial \lambda) \log L_{fU}(\lambda) | y_s, x_s, I_U, z_U]. \quad (3.28)$$

Another identity defines the relationship between the population likelihood information matrix and the sample likelihood information matrix.

Breckling, Chambers, Dorfman, Tam and Welsh (1994) and Chambers *et al.* (1998) consider applications of the MIP to complex survey data. In particular, Chambers *et al.* (1998) study the use of the MIP when only limited design information is available and not the full information entailed in z_U . The authors show examples where the use of the MIP is more efficient than the use of the sample likelihood $L_s(\theta, \gamma; y_s, x_s)$ defined by (3.19), which only uses the weights $\{w_i, i \in s\}$. The likelihood (3.28) can be extended to account for NMAR nonresponse but the application of this approach requires then knowledge of the population values of the variables explaining the response. The computation of the expectation in the right hand side of (3.29) may not be simple either, depending on the population model.

Remark 11. The use of the MIP method in the simulation set up of Section (3.1) requires knowledge of the covariates and stratification membership for units outside the sample. We didn't find a way of applying the method in this case without further assumptions on the joint distribution of the covariates and the design variables.

3.6.3 Empirical likelihood

In recent years there is a growing interest in the use of empirical likelihood (EL) methods for analyzing complex survey data. The EL method as originally proposed by Hartley and Rao (1968) in the survey sample context and by Owen (1988, 2001) combines the robustness of non-parametric methods with the effectiveness of the likelihood approach. Two other important advantages of this method are that it lends itself very naturally to the use of calibration equations and that it enables the construction of confidence intervals without the need for variance estimation.

Consider the model defined by (3.13) where for now we view the covariates as random, and denote $g_i = (y_i, x_i)'$. Under some regularity conditions, the vector parameter θ is the unique solution of the equation

$$E_p \left\{ \frac{\partial m(x; \theta)}{\partial \theta} [y - m(x; \theta)] \right\} = 0.$$

Let p_1, \dots, p_n be a set of probabilities corresponding to the observations (g_1, \dots, g_n) such that p_i is the 'jump' (probability mass) of the population cumulative distribution $F_p(g_i)$ at g_i . It is assumed that F_p has its support on the observed values such that

$$\sum_{i=1}^n p_i \frac{\partial m(x_i; \theta)}{\partial \theta} [y_i - m(x_i; \theta)] = 0. \quad (3.29)$$

Assuming independent observations, the EL of F_p is $L(F_p) = \prod_{i=1}^n p_i$. Notice that if p_i is a known function of some unknown parameters, $L(F_p)$ coincides with the standard parametric likelihood. The (nonparametric) EL estimators of the probabilities p_i are the solution $p_i^{(p)}$ of the maximization problem,

$$\max_{p_1, \dots, p_n} \prod_{i=1}^n p_i \text{ s.t. } p_i \geq 0, \sum_{i=1}^n p_i = 1, \quad (3.30)$$

yielding $p_i^{(p)} = 1/n, i = 1, \dots, n$. For the linear regression case, $m(x_i; \theta) = x_i' \beta$ and by substituting $p_i^{(p)}$ for p_i in (3.29) and solving the equations we obtain the EL estimator of β as $\hat{\beta}_{el} = \hat{\beta}_{OLS}$. When finite population means \bar{C}^U of variables C measured in the sample are known, they can be added to the maximization problem (3.30) by adding the calibration constraints $\sum_{i=1}^n p_i c_i = \bar{C}^U$. This additional information is expected to enhance the estimation of the p_i 's and hence the estimation of the unknown model parameters. See also Remark 12 below.

Suppose now that units are drawn to the sample (or respond) with unequal selection probabilities π_i . In this case it is common to replace the objective empirical likelihood $L(F_p) = \prod_{i=1}^n p_i$ by the pseudo empirical likelihood $L_{pl}(F_p) = \prod_{i=1}^n p_i^{w_i}$, where, as before, $w_i = 1/\pi_i$. Notice that $\log L_{pl}(F_p) = \sum_{i=1}^n w_i \log(p_i)$ is the H-T estimator of $\log L_{pop}(F_p) = \sum_{i=1}^N \log p_i$. The pseudo EL estimators of the p_i 's solve the maximization problem,

$$\max_{p_1, \dots, p_n} \prod_{i=1}^n p_i^{w_i} \text{ s.t. } p_i \geq 0, \sum_{i=1}^n p_i = 1. \quad (3.31)$$

See, e.g., Chen and Sitter (1999). It is easy to verify that in the absence of benchmark constraints, the solution of (3.31) is $p_i^{(pel)} = w_i / \sum_{i=1}^n w_i$ and by substituting $p_i^{(pel)}$ for p_i in (3.29), $\hat{\beta}_{pel} = \hat{\beta}_{pw}$, the PW estimator (3.8).

The empirical likelihoods in (3.30) and (3.31) are with respect to the population distribution. Alternatively, one can obtain the EL estimator by defining the likelihood with respect to the sample distribution $f_s(g_i) = \Pr(I_i = 1 | g_i) f_p(g_i) / \Pr(I_i = 1)$, where by denoting $\tau_i = \Pr(I_i = 1 | g_i)$, $\Pr(I_i = 1) = \sum_{i=1}^n p_i \tau_i$. Following Kim (2009) and Chaudhuri, Handcock and Rendall (2010), the EL estimators of the probabilities p_i are obtained now as the solution of the maximization problem

$$\max_{p_1, \dots, p_n} \left[\sum_{i=1}^n \log(p_i \tau_i) - n \log \sum_{i=1}^n p_i \tau_i \right] \\ \text{s.t. } p_i \geq 0, \sum_{i=1}^n p_i = 1. \quad (3.32)$$

The solution of (3.32) is $p_i^{\text{sel}} = \tau_i^{-1} / \sum_{j=1}^n \tau_j^{-1}$ and by substituting in (3.29),

$$\hat{\beta}_{\text{sel}} = \left[\sum_{i=1}^n \tau_i^{-1} x_i x_i' \right]^{-1} \sum_{i=1}^n \tau_i^{-1} x_i y_i. \quad (3.33)$$

The estimator $\hat{\beta}_{\text{sel}}$ has the same form as the PW estimator $\hat{\beta}_{\text{pw}}$ in (3.8), but with the weights $\tau_i^{-1} = 1 / \Pr(i \in s | y_i, x_i)$ instead of the sampling weights w_i . In practice, one has to replace the probabilities τ_i by sample estimates $\hat{\tau}_i$. See Section 4.

Remark 12. The following possible enhancement to the estimation of the probabilities p_i was proposed to me by Dr. Jae Kim in a private communication. Assuming as before that $\Pr(i \in s | \pi_i, y_i, x_i) = \pi_i$, it follows that $\tau_i = \Pr(I_i = 1 | y_i, x_i) = E_p(\pi_i | y_i, x_i)$ and hence that $E_p[(\pi_i - \tau_i) | y_i, x_i] = 0$. This suggests adding calibration constraints of the form

$$\sum_{j=1}^n p_j (\pi_j - \hat{\tau}_j) k(y_j, x_j) = 0 \quad (3.34)$$

to enhance the estimation of the probabilities $\{p_i\}$ in (3.31), where $k(y_j, x_j) = k(g_j)$ is some function of the observed outcome and covariates. Examples for plausible functions for the case of a single covariate x are, $k(g_j) = y_j x_j$, $k(g_j) = y_j / x_j$ etc. The notable feature of the constraints (3.34) is that they do not require knowledge of population quantities like means of calibration variables, as is often assumed when advocating the EL approach for sample survey estimation. Clearly, when means \bar{C}^U of calibration variables are known, constraints of the form $\sum_{i=1}^n p_i c_i = \bar{C}^U$ may be added as well. See also Remark 14.

4. Empirical study

In this section I report the results of a simulation study aimed at assessing and comparing the performance of the methods discussed in Section 3. The simulation set up is described in Section 3.1 and we use $H = 5$ strata. The target parameters are the regression coefficients $\beta' = (\beta_0, \beta_1) = (2, 1)$ of the population expectation (3.1). The simulation study consists of generating 2,000 populations and samples (one sample from each population) and computing the estimators, variance estimators and confidence intervals listed below for each sample. The population size is 5,000 with approximate strata sizes $N_h = 363, 554, 842, 1,278, 1,963$. (The strata sizes are random). The sample size is $n = 300$ with $n_h = 60$ sampled units in each stratum. The sampling fractions are therefore highly variable across the strata.

We generated population values of a single discrete covariate x by first generating observations \tilde{x}_j from a *Gamma* distribution with mean 2 and variance 4, and then defining x_j to be the nearest integer to \tilde{x}_j if $\tilde{x}_j < 5$ and $x_j = 5$ otherwise. The covariates are therefore $x_j = (1, x_j)'$, with $x_j = 0, 1, \dots, 5$. The population covariates were generated once and held fixed for all the populations.

Figure 1 shows the population and sample *pdfs* of the outcome y for $x = 2, 3, 4, 5$.

As can be seen, the population and sample *pdfs* differ, indicating the informativeness of the sampling process. Notice also that the population *pdf* is not normal because the random coefficients ζ_j are not normal.

We study the performance of the various methods in terms of *bias*, *variance*, *variance estimation*, and *confidence interval* coverage. We assume for all the methods that the only available information are the observed outcomes and covariates (y_{hs}, x_{hs}) for every stratum h , the sample selection probabilities and the true strata sizes $\{N_h\}$. I believe that this is the practice in most real life applications.

4.1 Estimators considered

4.1.1 The OLS estimator $\hat{\beta}_{\text{ols}}$. The use of this estimator ignores the sampling process.

4.1.2 The estimator proposed by Feder (2011, see Section 3.2). Application of this approach is in four steps. *i*) fit a linear model with constant residual variance in each stratum, *ii*) impute the missing covariate values for the non-sampled units by sampling with replacement $(N_h - n_h)$ values from the n_h observed values in stratum h with probabilities $\tilde{p}_{hi} = (w_{hi} - 1) / \sum_{k=1}^{n_h} (w_{hk} - 1)$ on each draw, where the w_{hi} 's are the sampling weights when sampling from stratum h . *iii*) impute the missing y -values in each stratum by generating observations at random from the model fitted in Step *i*). *iv*) fit the linear regression model of y on x by using all the population data, with the missing values for the non-sampled units replaced by the imputed values. We denote the resulting estimator by $\hat{\beta}_f$.

4.1.3 The PW estimator $\hat{\beta}_{\text{pw}}$ (Equation 3.8).

4.1.4 The estimator $\hat{\beta}_{\text{mg}}$ proposed by Magee (1998, see Section 3.5). In our application we define $a_i(\alpha) = (x_i + 0.1)^\alpha$ and search for the optimal power α in the range $[-2, 2]$ minimizing the determinant of the asymptotic variance estimator (3.12).

4.1.5 The estimator $\hat{\beta}_q$ defined by (3.16). For the present study we do not assume any parametric model for the expectation $E_s(w_i | x_i)$ in the denominator of q_i and estimate $\hat{E}_s(w_i | x_i) = \bar{w}_s(x_i)$, the mean of the observed sampling weights for units with $x = x_i$.

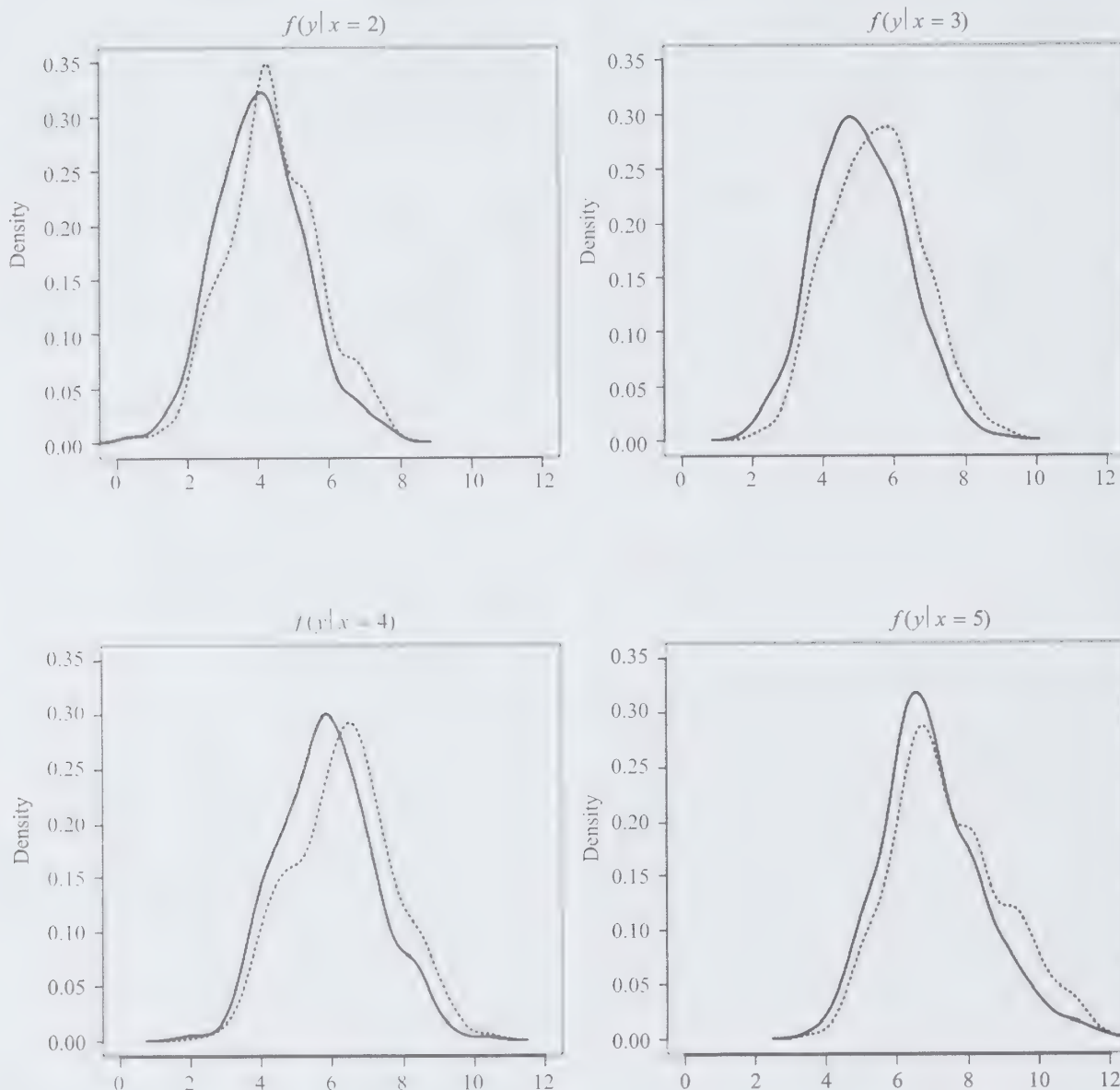


Figure 1 Population *pdf* (solid line) and sample *pdf* (dashed line) of $y|x$

4.1.6 The modified q -weighted estimator $\hat{\beta}_{\text{mg-}q}$ defined by (3.17). The weights \hat{q}_i are obtained as in 4.1.5 and the functions $a_{i,q}(\alpha)$ as in 4.1.4.

4.1.7 Estimators derived by maximization of the sample likelihood (3.19). The use of this approach requires specifying the population *pdf* and the expectation $E_s(w_i | y_i, x_i)$. The unknown population model parameters are $\theta' = (\beta', \sigma^2)$ and we assume $f_p(y_i | x_i; \theta) = N(x_i' \beta, \sigma^2)$, which as noted before and illustrated in Figure 1 is not the correct *pdf* since the random coefficients ζ_i are not normal (see Section 3.1). We estimated $E_s(w_i | y_i, x_i; \gamma)$ nonparametrically and set up the likelihood as follows:

Let s_{x_i} define the sample of units with $x = x_i$ of size m_{x_i} . We first divided the sample into $c(x_i)$ homogeneous clusters based on the ascending values of the outcome y using the R function “hclust”. The $c(x_i)$ ’s are between 1 and 7, depending on the sample size m_{x_i} (one cluster if $m_{x_i} \leq 10$, 2 clusters if $m_{x_i} \leq 20$, ..., 7 clusters if $m_{x_i} \geq 70$). Denote by $b_{x_i,k}$ the midpoint between the highest y -value in cluster k and the lowest y -value in cluster $(k+1)$, $k = 1, \dots, c(x_i) - 1$, and define $b_{x_i,0} = -\infty$, $b_{x_i,c(x_i)} = +\infty$. For $b_{x_i,k-1} \leq y \leq b_{x_i,k}$ we estimated $E_s(w_i | y_i, x_i)$ by the mean $\bar{w}_s(y, x_i) = \bar{w}_k(x_i)$ of the sampling weights of units with y -values in the same interval. Substituting $E_s(w_i | y_i, x_i) = \bar{w}_s(y, x_i)$ in (3.19) defines the sample likelihood used for the present simulation study as,

$$L_s(\theta; y_s, x_s) = \prod_{i \in s} \frac{f_p(y_i | x_i; \theta) / \bar{w}_s(y_i, x_i)}{\sum_{k=1}^{c(x_i)} [F_p(b_{k, x_i}) - F_p(b_{k-1, x_i})] / \bar{w}_k(x_i)}, \quad (4.1)$$

where $F_p(b_{k, x_i}) = \int_{-\infty}^{b_k} f_p(y | x_i; \theta) dy$ (the CDF of the assumed normal pdf).

The approximation (4.1) is similar to the approximation (3.20) proposed for the case where both x and y are discrete.

Remark 13. In order to facilitate the numerical optimizations used for the computation of the estimators $\hat{\beta}_{mg}$, $\hat{\beta}_{mg-q}$ and the maximum likelihood estimators in (4.1), we transformed the minimization problem $\min\{f(\theta); \theta \in (a, b)\}$ to $\min\{f[g(\eta)]; \eta \in (-\infty, \infty)\}$ with the function $g(\eta)$ defined as $g(\eta) = [(b-a)\tan^{-1}(\eta)] / \pi + 0.5(a+b)$. Notice that every $\theta \in (a, b)$ has an image $\eta \in R$; $g(\eta) = \theta$, and $\arg \min\{f(\theta); \theta \in (a, b)\} = g(\eta_0)$ where $\eta_0 = \arg \min f[g(\eta)]$.

We used the R function *nlm* for the numerical optimization, with the PW estimates as starting values. To prevent numerical overflows of the optimized function by evaluation of exponentials of large numbers, the maximization was limited to the intervals $\{\min[0.5\hat{\beta}_{pw}, \hat{\beta}_{pw} - 3\hat{se}(\hat{\beta}_{pw})], \max[1.5\hat{\beta}_{pw}, \hat{\beta}_{pw} + 3\hat{se}(\hat{\beta}_{pw})]\}$ for β , and $[0.5\hat{\sigma}_{pw}, 1.5\hat{\sigma}_{pw}]$ for σ .

4.1.8 The empirical likelihood estimator $\hat{\beta}_{sel}$ defined by (3.33). The computation of this estimator requires estimating the probabilities $\tau_i = \Pr(I_i = 1 | y_i, x_i) = 1/E_s(w_i | y_i, x_i)$, and we use the estimator $\hat{E}_s(w_i | y_i, x_i) = \bar{w}_{s,k}(y, x_i)$ used for defining the likelihood (4.1), such that $\hat{\tau}_i = 1 / \bar{w}_k(y, x_i)$.

4.2 Variance estimation

We applied three approaches for variance estimation. The first approach estimates the randomization variance, the second approach estimates the variance under the sample model, while the third approach uses the nonparametric bootstrap method, which likewise estimates the variance under the sample model.

Consider first the estimators defined by 4.1.1, 4.1.3 – 4.1.6 and 4.1.8 in Section 4.1. All these estimators can be written in the generic form,

$$\hat{\beta}_t = \left[\sum_{i=1}^n w_i t_i x_i x_i' \right]^{-1} \sum_{i=1}^n w_i t_i x_i y_i = [X_s' W_s T_s X_s]^{-1} \sum_{i=1}^n w_i t_i x_i y_i, \quad (4.2)$$

where $X_s' = [x_1, \dots, x_n]$, $W_s = \text{diag}[w_1, \dots, w_n]$ is the diagonal matrix with the sampling weights on the main diagonal and $T_s = \text{diag}[t_1, \dots, t_n]$, with the t_i 's defined by the estimators. For $\hat{\beta}_{ols}$ $t_i = 1/w_i$, for $\hat{\beta}_{sel}$ $t_i = w_i^{-1} \hat{\tau}_i^{-1}$ and so forth. The randomization variance of these estimators is estimated as,

$$\begin{aligned} \text{V}\hat{\text{ar}}_r(\hat{\beta}_t) &= [X_s' W_s T_s X_s]^{-1} [\text{V}\hat{\text{ar}}_r \sum_{i=1}^n w_i t_i x_i e_{it}] [X_s' W_s T_s X_s]^{-1}, \quad (4.3) \end{aligned}$$

where $e_{it} = (y_i - x_i' B)$ and B is the census estimator. Using the double index (hj) to define the j^{th} unit in the sample s_h of size n_h drawn from stratum h , we estimated

$$\begin{aligned} \text{V}\hat{\text{ar}}_r \left[\sum_{i=1}^n w_i t_i x_i e_{it} \right] &= \sum_{h=1}^5 \text{V}\hat{\text{ar}} \left(\sum_{j=1}^{n_h} w_{hj} \tilde{e}_{hj,t} \right) \\ &= \sum_{h=1}^5 \frac{n_h}{(n_h - 1)} \sum_{j=1}^{n_h} (w_{hj} \tilde{e}_{hj,t} - \bar{e}_{h,t})(w_{hj} \tilde{e}_{hj,t} - \bar{e}_{h,t})', \quad (4.4) \end{aligned}$$

where $\tilde{e}_{hj,t} = t_{hj} x_{hj} (y_{hj} - x_{hj}' \hat{\beta}_t)$ and

$$\bar{e}_{h,t} = \frac{1}{n_h} \sum_{j=1}^{n_h} w_{hj} \tilde{e}_{hj,t},$$

assuming with replacement sampling within the strata.

A variance estimator under the sample model which accounts for possible heteroscedasticity is obtained as,

$$\begin{aligned} \text{V}\hat{\text{ar}}_{sm}(\hat{\beta}_t) &= [X_s' W_s T_s X_s]^{-1} \left[\sum_{i \in s} w_i^2 t_i^2 \hat{e}_{it}^2 x_i x_i' \right] [X_s' W_s T_s X_s]^{-1}, \quad (4.5) \end{aligned}$$

where $\hat{e}_{it} = (y_i - x_i' \hat{\beta}_t)$. Randomization and sample model variance estimators for the estimator in 4.1.2 are developed by Feder (2011). For the maximum likelihood estimator under the sample model with the likelihood defined by (4.1) we only estimate the variance under the sample model using the inverse information matrix.

Finally, bootstrap variance estimators for all the estimators are obtained by sampling with replacement n units from the original sample and re-estimating each of the estimators using the same computations as for the original sample. Repeating the same process independently B times, the bootstrap variance estimator is,

$$\begin{aligned} \text{V}\hat{\text{ar}}_{BS}(\hat{\beta}) &= \frac{1}{B} \sum_{b=1}^B (\hat{\beta}^{(b)} - \bar{\hat{\beta}})(\hat{\beta}^{(b)} - \bar{\hat{\beta}})', \\ \bar{\hat{\beta}} &= \frac{1}{B} \sum_{b=1}^B \hat{\beta}^{(b)}, \quad (4.6) \end{aligned}$$

where $\hat{\beta}$ represents any of the estimators defined by 4.1.1 – 4.1.8 and $\hat{\beta}^{(b)}$ is the corresponding estimator computed for bootstrap sample b , $b = 1, \dots, B$.

4.3 Computation of confidence intervals

We consider two approaches of $(1 - \alpha)$ level confidence interval (C.I.) computation. The first approach is the standard C.I.,

$$\hat{\beta}_k \pm Z_{1-\frac{\alpha}{2}} \hat{s.e.}(\hat{\beta}_k), k = 0, 1,$$

where $\hat{\beta}_k$ stands for any of the estimators considered and $\hat{s.e.}(\hat{\beta}_k)$ is the corresponding estimator of the standard error as obtained by one of the methods listed before. The second, “basic bootstrap” approach uses the quantiles $bs(k, \tilde{\alpha})$ of the bootstrap estimators $\hat{\beta}_k^{(b)}$ to compute the C.I.

$$\left[2\hat{\beta}_k - bs\left(k, 1 - \frac{\alpha}{2}\right), 2\hat{\beta}_k - bs\left(k, \frac{\alpha}{2}\right) \right], k = 1, 2.$$

We tried also the use of the “studentized bootstrap method” but the coverage rates were not better with any of the estimators $\hat{\beta}_k$. See Remark 14 below.

4.4 Simulation results

Table 1 shows the empirical means of the estimates listed in Section 4.1 over the 2,000 populations and samples and the corresponding empirical standard errors (S.E.). Also shown are the square roots of the means of the variance estimates as obtained when estimating the randomization variance (“Ran.”) and when estimating the variance under the sample model (“S.M.”). Because of computing time limitations, the results for the bootstrap variance estimators (“BS”) are based on 300 bootstrap samples drawn from each of 500 original samples. These numbers of original and bootstrap samples were found to produce stable variance estimators.

As expected, given the use of an informative sampling scheme, the OLS estimator has a relatively large bias of 12% (5%) when estimating the intercept (slope). All the other estimators are virtually unbiased, except for $\hat{\beta}_{mle}$, which has bias of 2% and 1.5%. The almost unbiasedness of the EL estimator $\hat{\beta}_{sel}$ is particularly encouraging given the somewhat crude nonparametric estimation of the probabilities $\tau_i = \Pr(i \in s | y_i, x_i)$. Notice also that this estimator has similar empirical S.E. to those of the PW estimator. The small (but statistically significant) bias of $\hat{\beta}_{mle}$ is explained by the fact that we assume a normal distribution under the population model, which as noted and illustrated before is incorrect.

Regarding precision, the OLS estimator has the smallest S.E. but $\hat{\beta}_f$ has almost the same S.E. (and is unbiased). This is explained by the fact that this estimator uses additional stratification information, not used by the other estimators. Note that $\hat{\beta}_{mg}$, $\hat{\beta}_{mg-q}$ and particularly $\hat{\beta}_q$ outperform $\hat{\beta}_{pw}$, but $\hat{\beta}_{mg-q}$ does not improve over $\hat{\beta}_q$.

Remark 14. Following my presentation of this paper at the 2011 Statistics Canada symposium, Jean-Francois Beaumont suggested to replace the weights $\hat{\tau}_i^{-1}$ used for the computation of $\hat{\beta}_{sel}$ by the weights $\hat{\tau}_i^{-1} / E_s(\hat{\tau}_i^{-1})$, so as to account for the net sampling effects on the conditional pdf $f(y | x)$, similarly to the use of the q -weights in $\hat{\beta}_q$. Notice that whereas the sampling weights w_i may depend on y , x and possibly other variables, the weights $\hat{\tau}_i^{-1}$ only depend on y and x . Application of this idea did not affect the bias but the empirical S.E. of the modified estimators are 0.151 and 0.053, smaller than the S.E. of $\hat{\beta}_{sel}$ and similar to the S.E. of $\hat{\beta}_q$.

Looking at the performance of the variance estimators, the first remarkable outcome is that the randomization and sample model variance estimators (Equations 4.4 and 4.5) are very similar for every estimator of the regression coefficients, even though they are computed very differently. For $\hat{\beta}_{ols}$, $\hat{\beta}_{pw}$ and $\hat{\beta}_q$ the variance estimators are almost unbiased but for the other estimators the variance estimators under-estimate the true variance. This is explained by the fact that these variance estimators ignore some of the operations involved in the computation of the estimated regression coefficients. Thus, in the case of the estimators $\hat{\beta}_{mg}$ and $\hat{\beta}_{mg-q}$ the variance estimators do not account for the choice of the optimal weights $a_i(\alpha)$, in the case of $\hat{\beta}_f$ the variance estimator does not account for the random imputation of the vectors (y_i, x_i) for $i \in U - s$, and in the case of $\hat{\beta}_{mle}$ and $\hat{\beta}_{sel}$ the variance estimators do not account for the estimation of the probabilities $\Pr(i \in s | y_i, x_i)$. This under-estimation of the variance is corrected in almost all cases by use of the bootstrap method, see, in particular, the estimation of the variances of $\hat{\beta}_f$, $\hat{\beta}_{mle}$ and $\hat{\beta}_{sel}$.

Figure 2 shows the empirical coverage rates of $(1 - \alpha)$ -level confidence intervals (C.I.) for $\alpha = 0.10, 0.05, 0.01$, as obtained when applying the standard C.I. with the standard errors estimated by the BS method, and when using the basic bootstrap method. The figures in the horizontal axis are the nominal levels

The coverage rates are almost always below the nominal levels but the under-coverage in the case of the standard C.I. is generally less than 4%. The two exceptions are when basing the confidence intervals on the OLS estimators (large under-coverage) and the mle estimator of the slope (under-coverage of 7% at the 90% nominal level), which is explained by the bias of these estimators. The under-coverage percentages when using the basic bootstrap method are generally slightly larger, except for the under-coverage of the C.I. for the intercept based on $\hat{\beta}_{sel}$, which is more pronounced.

Table 1
Means, standard errors (S.E.) and square roots of means of variance estimates. Population model: $E_p(y_j) = 2 + 1 \times x_j$, $\text{Var}_p(y_j) = (1 + 0.2x_j)^2 V_j + 1$

Method	Intercept- $\hat{\beta}_0$					Slope- $\hat{\beta}_1$				
	Mean Est.	Emp. S.E.	Ran.	S.M.	BS	Mean Est.	Emp. S.E.	Ran.	S.M.	BS
$\hat{\beta}_{ols}$	2.251	0.133	0.135	0.139	0.140	1.046	0.048	0.048	0.049	0.049
$\hat{\beta}_f$	2.006	0.133	0.126	0.126	0.135	0.999	0.051	0.041	0.041	0.052
$\hat{\beta}_{pw}$	2.008	0.166	0.167	0.169	0.157	0.998	0.059	0.055	0.055	0.056
$\hat{\beta}_{mg}$	2.017	0.158	0.154	0.156	0.154	0.995	0.056	0.050	0.050	0.055
$\hat{\beta}_q$	2.011	0.153	0.157	0.159	0.147	0.999	0.054	0.051	0.051	0.052
$\hat{\beta}_{mg-q}$	2.020	0.156	0.152	0.154	0.153	0.996	0.055	0.049	0.050	0.054
$\hat{\beta}_{mle}$	1.960	0.159	----	0.143	0.152	1.026	0.054	----	0.046	0.053
$\hat{\beta}_{sel}$	2.031	0.164	0.143	0.143	0.159	0.995	0.058	0.049	0.049	0.057

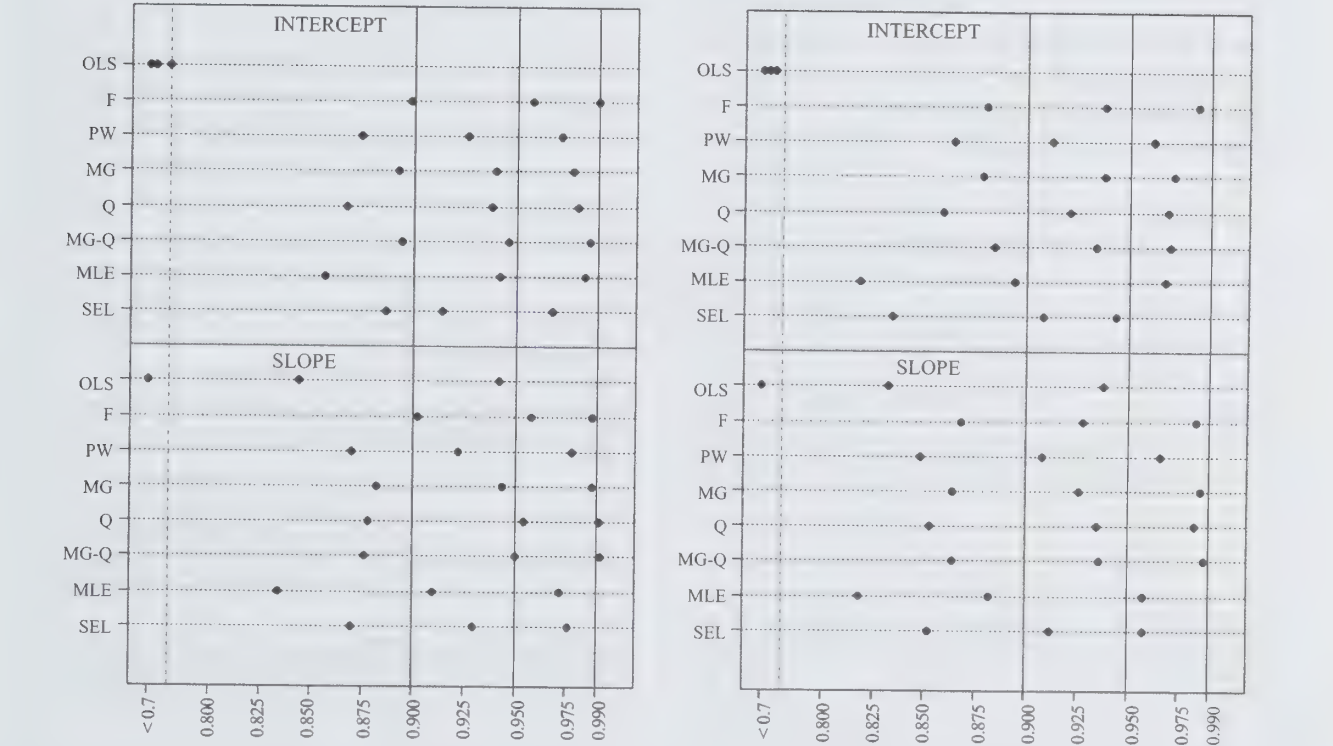


Figure 2 Coverage rates of standard (left) and BS (right) confidence intervals

Remark 15. We computed also the standard C.I. with the S.E. estimated under the randomization distribution (Equation 4.4) and under the sample model (Equation 4.5), but except in the case of the estimators $\hat{\beta}_{pw}$ and $\hat{\beta}_q$, the under-coverage of these intervals was somewhat higher than the coverage rates in Figure 2 because of the underestimation of the true S.E. by these S.E. estimators discussed before. The same phenomenon was observed when using the “studentized bootstrap method” with these S.E. estimates, which again can be explained by the

underestimation of the true S.E.’s. The use of more advanced bootstrap C.I. such as double-bootstrap may correct this under-coverage.

5. Concluding remarks

In this article I discuss alternative procedures proposed in the literature to account for informative sampling and NMAR nonresponse when modeling survey data. The empirical study is restricted so far to the case of linear

regression and single-stage sampling, and an obvious extension would be to consider other models and cluster sampling. The present study illustrates the unbiasedness or approximate unbiasedness of all the point estimators considered, but the standard variance estimators underestimate the true variances in most cases since they fail to account for the extra operations involved in computing the corresponding point estimators. The bootstrap variance estimators produce much better variance estimators in these cases. The confidence intervals applied in the present study yield small under-coverage in most cases, but they should be improved, possibly by use of more advanced bootstrap techniques. Another important extension mentioned in the paper, which we have not investigated empirically so far is to incorporate sample based calibration constraints in the empirical likelihood method when based on the sample distribution.

We plan to apply the various methods to several real data sets. This would require the development of diagnostic procedures that would allow comparing the performance of the methods since unlike in a simulation study, the true distributions and model parameters are seldom known in real applications.

Acknowledgements

I am indebted to Dr. Moshe Feder for carrying out the empirical study and many helpful comments and suggestions. Thanks are due also to Dr. Pedro Silva for his constructive remarks on an earlier draft of the paper and three reviewers for their careful reading and comments in a short time period given to them. This study is funded by a UK ESRC grant No. RES-062-23-2316.

References

- Binder, D.A. (1983). On the variances of asymptotically normal estimators from complex surveys. *International Statistical Review*, 51, 279-292.
- Binder, D., and Roberts, G. (2009). Design and model based inference for model parameters. In *Handbook of Statistics 29B; Sample Surveys: Inference and Analysis*, (Eds., D. Pfeffermann and C.R. Rao). Amsterdam: North Holland, 33-54.
- Breckling, J.U., Chambers, R.L., Dorfman, A.H., Tam, S.M. and Welsh, A.H. (1994). Maximum likelihood inference from sample survey data. *International Statistical Review*, 62, 349-363.
- Brick, J.M., and Montaquila, J.M. (2009). Nonresponse and weighting. In *Handbook of Statistics 29A; Sample Surveys: Inference and Analysis*, (Eds., D. Pfeffermann and C.R. Rao). Amsterdam: North Holland, 163-185.
- Chambers, R.L., Dorfman, A.H. and Wang, S. (1998). Limited information likelihood analysis of survey data. *Journal of the Royal Statistical Society, Series B*, 60, 397-411.
- Chambers, R.L., and Skinner, C.J. (2003, Eds.). *Analysis of survey data*. New York: John Wiley & Sons, Inc.
- Chambless, L.E., and Boyle, K.E. (1985). Maximum likelihood methods for complex sample data: Logistic, regression and discrete proportional hazards models. *Communication in Statistics-Theory and Methods*, 14, 1377-1392.
- Chang, T., and Kott, P.S. (2008). Using calibration weighting to adjust for nonresponse under a plausible model. *Biometrika*, 95, 555-571.
- Chen, J., and Sitter, R.R. (1999). A pseudo empirical likelihood approach to the effective use of auxiliary information in complex surveys. *Statistica sinica*, 9, 385-406.
- Chaudhuri, S., Handcock, M.S. and Rendall, M.S. (2010). A conditional empirical likelihood approach to combine sampling design and population level information. Technical report No. 3/2010, National University of Singapore, Singapore, 117546.
- DeMets, D., and Halperin, M. (1977). Estimation of simple regression coefficients in samples arising from sub-sampling procedures. *Biometrics*, 33, 47-56.
- DuMouchel, W.H., and Duncan, G.L. (1983). Using sample survey weights in multiple regression analysis of stratified samples. *Journal of the American Statistical Association*, 78, 535-543.
- Feder, M. (2011). Fitting Regression Models to Complex Survey Data- Gelman's Estimator Revisited. In Proceedings of the ISI meeting, Ireland, (www.isi2011.ie).
- Francisco, C.A., and Fuller, W.A. (1991). Quantile estimation with a complex survey design. *The Annals of Statistics*, 19, 454-469.
- Fuller, W.A. (1975). Regression analysis for sample surveys. *Sankhyā, Series C*, 37, 117-132.
- Fuller, W.A. (2002). Regression estimation for survey samples. *Survey Methodology*, 28, 5-23.
- Gelman, A., Carlin, J.B., Stern, H.S. and Rubin, D.B. (2003). *Bayesian Data Analysis*, second edition. London: CRC Press.
- Gelman, A. (2007). Struggles with survey weighting and regression modeling (with discussion). *Statistical Science*, 22, 153-164.
- Godambe, V.P., and Thompson, M.E. (1986). Parameters of superpopulation and survey population: Their relationships and estimation. *International Statistical Review*, 54, 127-138.
- Godambe, V.P., and Thompson, M.E. (2009). Estimating functions and survey sampling. In *Handbook of Statistics 29B; Sample Surveys: Inference and Analysis*, (Eds., D. Pfeffermann and C.R. Rao). Amsterdam: North Holland, 83-101.
- Goldstein, H. (1986). Multi-level mixed linear model analysis using iterative generalized least squares. *Biometrika*, 73, 43-56.
- Häjek, J. (1971). Comments on a paper by D. Basu. In *Foundations of Statistical Inference*, (Eds., V.P. Godambe and D.A. Sprott). Toronto: Holt, Rinehart and Winston.
- Hartley, H.O., and Rao, J.N.K. (1968). A new estimation theory for sample surveys. *Biometrika*, 55, 547-557.

- Holt, D., Smith, T.M.F. and Winter, P.D. (1980). Regression analysis of data from complex surveys. *Journal of the Royal Statistical Society, Series A*, 143, 474-487.
- Jewell, N.P. (1985). Least squares regression with data arising from stratified samples of the dependent variable. *Biometrika*, 72, 11-21.
- Kasprzyk, D., Duncan, G.J., Kalton, G. and Singh, M.P. (1989, Eds.). *Panel Surveys*. New York: John Wiley & Sons, Inc.
- Kim, J.K. (2009). Calibration estimation using empirical likelihood in survey sampling. *Statistica Sinica*, 19, 145-157.
- Kott, P.S. (2009). Calibration Weighting: Combining Probability Samples and Linear Prediction Models. In *Handbook of Statistics 29B; Sample Surveys: Inference and Analysis*, (Eds., D. Pfeffermann and C.R. Rao). Amsterdam: North Holland, 55-82.
- Krieger, A.M., and Pfeffermann D. (1997). Testing of distribution functions from complex sample surveys. *Journal of Official Statistics*, 13, 123-142.
- Little, R.J.A. (1982). Models for non-response in sample surveys. *Journal of the American Statistical Association*, 77, 237-249.
- Little, R.J.A. (2004). To model or not to model? Competing modes of inference for finite population sampling. *Journal of the American Statistical Association*, 99, 546-556.
- Magee, L. (1998). Improving survey-weighted least squares regression. *Journal of the Royal Statistical Society, Series B*, 60, 115-126.
- Nathan, G., and Holt, D. (1980). The effect of survey design on regression analysis. *Journal of the Royal Statistical Society, Series B*, 42, 377-386.
- Orchard, T., and Woodbury, M.A. (1972). A missing information principle: Theory and application. *Proceedings of the 6th Berkeley Symposium on Mathematical Statistics and Probability*, 1, 697-715.
- Owen, A.B. (1988). Empirical likelihood ratio confidence intervals for a single functional. *Biometrika*, 75, 237-249.
- Owen, A.B. (2001). *Empirical likelihood*. New York: Chapman & Hall.
- Pfeffermann, D., and Holmes, D. (1985). Robustness consideration in the choice of method of inference for regression analysis of survey data. *Journal of the Royal Statistical Society, Series A*, 148, 268-278.
- Pfeffermann, D., and Smith, T.M.F. (1985). Regression models for grouped populations in cross-section surveys. *International Statistical Review*, 53, 37-59.
- Pfeffermann, D. (1993). The role of sampling weights when modeling survey data. *International Statistical Review*, 61, 317-337.
- Pfeffermann, D. (1996). The use of sampling weights for survey data analysis. *Statistical Methods in Medical Research*, 5, 239-261.
- Pfeffermann, D., Krieger, A.M. and Rinott, Y. (1998a). Parametric distributions of complex survey data under informative probability sampling. *Statistica Sinica*, 8, 1087-1114.
- Pfeffermann, D., Skinner, C.J., Holmes, D.J., Goldstein, H. and Rasbash, J. (1998b). Weighting for unequal selection probabilities in multi-level models (with discussion). *Journal of the Royal Statistical Society, Series B*, 60, 23-76.
- Pfeffermann, D., and Sverchkov, M. (1999). Parametric and semi-parametric estimation of regression models fitted to survey data. *Sankhyā*, 61, 166-186.
- Pfeffermann, D., and Sverchkov, M. (2003). Fitting generalized linear models under informative probability sampling. In *Analysis of Survey Data*, (Eds., R.L. Chambers and C.J. Skinner). New York: John Wiley & Sons, Inc., 175-195.
- Pfeffermann, D., Moura, F.A.S. and Nascimento-Silva, P.L. (2006). Multilevel modeling under informative sampling. *Biometrika*, 93, 943-959.
- Pfeffermann, D., and Sverchkov, M. (2007). Small area estimation under informative probability sampling of areas and within the selected areas. *Journal of the American Statistical Association*, 102, 1427-1439.
- Pfeffermann, D., and Sverchkov, M. (2009). Inference under Informative Sampling. In *Handbook of Statistics 29B; Sample Surveys: Inference and Analysis*, (Eds., D. Pfeffermann and C.R. Rao). Amsterdam: North Holland, 455-487.
- Pfeffermann, D., and Landsman, V. (2011). Are private schools better than public schools? Appraisal for Ireland by methods for observational studies. *The Annals of Applied Statistics*, 5, 1726-1751.
- Pfeffermann, D., and Sikov, N. (2011). Imputation and estimation under nonignorable nonresponse in household surveys with missing covariate information. *Journal of Official Statistics*, 27, 181-209.
- Rubin, D.B. (1976). Inference and missing data. *Biometrika*, 63, 605-614.
- Rubin, D.B. (1985). The use of propensity scores in applied Bayesian inference. In *Bayesian Statistics 2*, (Eds., J.M. Bernardo, M.H. Degroot, D.V. Lindley and A.F.M. Smith), Elsevier Science Publishers B.V., 463-472.
- Särndal, C.-E., and Wright, R. (1984). Cosmetic form of estimators in survey sampling. *Scandinavian Journal of Statistics*, 11, 146-156.
- Scott, A.J., and Holt, D. (1982). The effect of two-stage sampling on ordinary least squares. *Journal of the American Statistical Association*, 77, 848-854.
- Scott, A.J., and Wild, C.J. (2009). Population-based case-control studies. In *Handbook of Statistics 29B; Sample Surveys: Inference and Analysis*, (Eds., D. Pfeffermann and C.R. Rao). Amsterdam: North Holland, 431-453.
- Skinner, C.J., Holt, D. and Smith, T.M.F. (Eds.) (1989). *Analysis of complex surveys*. New York: John Wiley & Sons, Inc.
- Skinner, C.J. (1994). Sample models and weights. *Proceedings of the Section on Survey Research Methods*, 133-142.
- Smith, T.M.F. (1988). To weight or not to weight, that is the question. In *Bayesian Statistics 3*, (Eds., J.M. Bernardo, M.H. Degroot, D.V. Lindley and A.F.M. Smith), Oxford University Press, 437-451.

Sugden, R.A., and Smith, T.M.F. (1984). Ignorable and informative designs in survey sampling inference. *Biometrika*, 71, 495-506.

Sverchkov, M., and Pfeffermann, D. (2004). Prediction of finite population totals based on the sample distribution. *Survey Methodology*, 30, 79-92.

Wu, Y.Y., and Fuller, W.A. (2006). Estimation of regression coefficients with unequal probability samples. *Proceedings of the American Statistical Association, Section on Survey Research Methods*, 3892-3899.

A Bayesian analysis of small area probabilities under a constraint

Balgobin Nandram and Hasanjan Sayit¹

Abstract

In many sample surveys there are items requesting binary response (*e.g.*, obese, not obese) from a number of small areas. Inference is required about the probability for a positive response (*e.g.*, obese) in each area, the probability being the same for all individuals in each area and different across areas. Because of the sparseness of the data within areas, direct estimators are not reliable, and there is a need to use data from other areas to improve inference for a specific area. Essentially, *a priori* the areas are assumed to be similar, and a hierarchical Bayesian model, the standard beta-binomial model, is a natural choice. The innovation is that a practitioner may have much-needed additional prior information about a linear combination of the probabilities. For example, a weighted average of the probabilities is a parameter, and information can be elicited about this parameter, thereby making the Bayesian paradigm appropriate. We have modified the standard beta-binomial model for small areas to incorporate the prior information on the linear combination of the probabilities, which we call a constraint. Thus, there are three cases. The practitioner (a) does not specify a constraint, (b) specifies a constraint and the parameter completely, and (c) specifies a constraint and information which can be used to construct a prior distribution for the parameter. The griddy Gibbs sampler is used to fit the models. To illustrate our method, we use an example on obesity of children in the National Health and Nutrition Examination Survey in which the small areas are formed by crossing school (middle, high), ethnicity (white, black, Mexican) and gender (male, female). We use a simulation study to assess some of the statistical features of our method. We have shown that the gain in precision beyond (a) is in the order with (b) larger than (c).

Key Words: Accept-reject algorithm; Binomial distribution; Generalized beta distribution; Griddy Gibbs sampler; Simulation.

1. Introduction

It is a standard practice to use models to “borrow strength” in small area estimation (Rao 2003). Owing to the sparseness of the data in each area, direct estimates for small areas are typically not reliable. Our procedure allows a practitioner to incorporate prior information about a linear combination of binomial probabilities, one for each area. This is a constraint that we include as a weighted average of the area probabilities in the standard beta-binomial model. The weighted average can be assumed known or unknown. In the case when this value is unknown, we consider the scenario when there is some information which can be elicited from an expert in the form of prior distribution. This is different from standard practice in design based survey sampling in which auxiliary information is incorporated as in ratio and regression estimators (Cochran 1977). When the value can be specified exactly, there will be an increase in precision because prior information is incorporated into the model.

The beta-binomial model has been studied extensively. For example, Nandram and Sedransk (1993), Nandram (1998) and Nandram and Choi (2002) show how to do Bayesian predictive inference of finite population proportions of the small areas for binomial and multinomial data. These models assume that the binomial probabilities share a common effect, thereby permitting adaptive pooling of the

data from small areas (or clusters). However, it is possible to improve on these models further by including additional information using covariates via generalized linear models (*e.g.*, see Ghosh, Natarajan, Stroud and Carlin 1998). It is worth noting that none of these works propose ways to incorporate prior information about linear combination of model parameters. Substantial gains in precision are expected when such prior information is incorporated in small area models; see Silvapulle and Sen (2006) for a book-length discussion of constrained statistical inference. It is also worth noting that Lazar, Meeden and Nelson (2008) showed how to include constraints in nonparametric Bayesian approach via a Polya urn scheme to predictive distribution of finite population parameters.

Our procedure is related to external benchmarking which occurs when a pre-specified estimator is obtained from external sources, such as a different survey, a census, or other administrative records. In benchmarking one wants the parts to add up to the whole. For example, when surveys are conducted over time, there are typically monthly surveys and annual surveys which are of much better quality than the monthly surveys. When the monthly surveys are estimated such that these estimates add up to the annual survey totals, there is a protection against model failure and therefore improved estimates (*i.e.*, reduced bias and possibly an increase in precision). These problems are prevalent in the government agencies especially in employment and sales;

1. Balgobin Nandram and Hasanjan Sayit, Department of Mathematical Sciences, Worcester Polytechnic Institute, 100 Institute Road, Worcester, MA 01609-2280. E-mail: balnan@wpi.edu, hs7@wpi.edu.

see Hillmer and Trabelsi (1987) for an example on retail sales of hardware stores from the U.S. Census Bureau.

Prior information from external benchmarking will lead to improved precision but can produce severely biased estimators as well. This will depend on how different the current survey is from the prior ones. Nandram, Toto and Choi (2011) applied external benchmarking to estimate the finite population means of small areas. The constraint is the finite population mean for the entire population is a prespecified value which again can be obtained from a prior survey, census or administrative records. In our current work we are not incorporating information about a linear combination of the finite population values, but rather we are inputting information about a linear combination of the superpopulation parameters (in this case binomial probabilities).

We consider the problem in which binomial counts are obtained from similar small areas, and inference is required about the binomial probabilities. In the conclusion, we discuss how to extend our method to obtain the predictive distribution of finite population proportions. The standard beta-binomial model may be inadequate, and additional prior information must be incorporated. Our thesis is that there is an increase in precision over the standard beta-binomial small area model when prior information about the weighted average of the probabilities (*e.g.*, average of the probabilities) is incorporated. That is, we incorporate prior information about a linear combination of binomial probabilities (a weighted average). The weights can be proportional to population sizes, and under proportional allocation they can be proportional to the sample sizes themselves. The purpose of incorporating prior information about the binomial probabilities is to increase precision, and at the same time one needs to control the bias.

It is much easier for a survey practitioner to specify the value of the overall probability rather than the individual area probabilities. That is, the overall probability can be specified with relatively much less error than the individual probabilities. Of course, one can specify the overall probability using prior information (a prior survey, census or administrative records), and so the specification of the overall probability will depend on the quality of the prior information. Thus, the problem falls naturally within the Bayesian paradigm because we are incorporating prior information about a parameter via a distribution. Thus, there will be gains in precision because of the extra information. However, a practitioner can still proceed when there is no prior information. One can use the ratio of the total success and total sample size over areas to form a reasonable specification of the overall probability which is typically not of interest. This estimate will have much higher precision than the one for individual areas. There will still be gain in

precision, but clearly such gain is due to using the current data (double use) and the constraint.

One example of a survey in which reliable information can be obtained to perform the benchmarking is the National Health Interview Survey (NHIS) which is conducted annually by the National Center for Health Statistics to assess an aspect of Health of the U.S. population. This is a population-based survey and there are many health indicators of interest; one of these indicators is the number of doctor visits made in the past two weeks, and an informative quantity is the proportion of people who made at least one doctor visit last year (*e.g.*, Nandram and Choi 2002). These proportions are useful for small domains formed by crossing age, race and sex for a particular state last year. Because the estimates over a state change very slowly over the previous years, the overall estimate from the year immediately preceding last year can be used as a reliable benchmark for last year. If a reliable estimate cannot be obtained for the benchmark, one can construct an informative prior distribution for it. For example, one can use the method of moments to equate the sample mean and sample variance of the overall estimates for the past few years to the mean and variance of a beta distribution to get a beta prior distribution. In either case, our procedure can be applied.

The plan of this paper is as follows. In Section 2 we describe the methodology. Specifically, we describe the standard beta-binomial model, and we develop two additional models to incorporate the extra information using appropriate prior distributions. We also describe posterior inference and how to perform the nonstandard computations. In Section 3 we describe an illustrative example on obesity, and a simulation study to assess empirically the statistical properties of our models. Section 4 has concluding remarks. We also discuss how to do Bayesian predictive inference for finite population proportions. While we discuss binary data, we also show how one can extend our method to polychotomous data.

2. Methodology

We show how to incorporate the constraint into the beta-binomial model in two ways, thereby providing a set of alternative models. In Section 2.1 we describe the models and in Section 2.2 we describe posterior inference. We attempt to explain what the constraint does to the estimates of the probabilities using an approximation. In Section 2.3 we describe the computation, and we describe a new algorithm as well.

2.1 Models

We assume that binary data are available from ℓ small areas, and we assume that the probability that an individual

responds in the i^{th} area is π_i , $i = 1, \dots, \ell$. Let n_i be the number of individuals sampled from the i^{th} area, $i = 1, \dots, \ell$. Also let s_i denote the number of individuals with the characteristic and $f_i = n_i - s_i$ be the number of individuals without the characteristic in the i^{th} area, $i = 1, 2, \dots, \ell$. Then the standard beta-binomial hierarchical Bayesian model is

$$s_i | \pi_i \sim \text{Binomial}(n_i, \pi_i), \quad (1)$$

$$\pi_i | \mu, \tau \sim \text{Beta}\{\mu\tau, (1-\mu)\tau\}, \quad i = 1, \dots, \ell \quad (2)$$

and

$$p(\mu, \tau) = \frac{1}{(1+\tau)^2}, \quad 0 < \mu < 1, \tau \geq 0. \quad (3)$$

We use a shrinkage prior for τ because it is proper and noninformative, and there are no conjugate priors. Priors of the form $p(\tau) \propto 1/\tau$ are discouraged; see, for example, Gelman (2006). Other alternatives are half Cauchy densities and gamma densities (one would need to specify the hyperparameters). Henceforth, we will call the model specified by (1), (2) and (3) the unrestricted (UR) model or Model 1.

We next describe the restricted model, which is an extension of the unrestricted model. We obtain a simple linear combination of the binomial probabilities. Letting $\tilde{\pi}_i = s_i/n_i$ and

$$\omega_i = \frac{n_i}{\sum_{i=1}^{\ell} n_i}, \quad i = 1, \dots, \ell,$$

we have

$$\frac{\sum_{i=1}^{\ell} s_i}{\sum_{i=1}^{\ell} n_i} = \sum_{i=1}^{\ell} \omega_i \tilde{\pi}_i.$$

Thus, taking the π_i unknown, the linear combination is $\sum_{i=1}^{\ell} \omega_i \pi_i$.

Therefore, we need to make an adjustment in (2) to incorporate the restriction, $\sum_{i=1}^{\ell} \omega_i \pi_i = \theta$ conditional on θ . We do so by introducing the variable $\phi = \sum_{i=1}^{\ell} \omega_i \pi_i - \theta$; so that the restriction is equivalent to $\phi = 0$. Now one of the variables, π_i , $i = 1, \dots, \ell$, is redundant. It is worth noting that one can choose any one of π_1, \dots, π_{ℓ} , and without loss of generality and for ease of exposition, we choose π_{ℓ} . Thus, to incorporate the restriction, we transform π_{ℓ} to $\phi = \sum_{i=1}^{\ell} \omega_i \pi_i - \theta$, keeping $\pi_1, \dots, \pi_{\ell-1}$ untransformed, and we let $\pi_{(\ell)} = (\pi_1, \dots, \pi_{\ell-1})'$.

As the jacobian is $1/\omega_{\ell}$,

$$p(\pi_{(\ell)}, \phi | \mu, \tau, \theta) =$$

$$\frac{1}{\omega_{\ell}} \prod_{i=1}^{\ell-1} \frac{\pi_i^{\mu\tau-1} (1-\pi_i)^{(1-\mu)\tau-1}}{B\{\mu\tau, (1-\mu)\tau\}} \times \frac{\left[\frac{\phi + \theta - \sum_{i=1}^{\ell-1} \omega_i \pi_i}{\omega_{\ell}} \right]^{\mu\tau-1} \left[1 - \frac{\phi + \theta - \sum_{i=1}^{\ell-1} \omega_i \pi_i}{\omega_{\ell}} \right]^{(1-\mu)\tau-1}}{B\{\mu\tau, (1-\mu)\tau\}}, \quad (4)$$

where

$$0 < \pi_i < 1, \quad i = 1, \dots, \ell,$$

$$0 < \mu < 1, \tau > 0, \phi + \theta - \omega_{\ell} \leq \sum_{i=1}^{\ell-1} \omega_i \pi_i \leq \phi + \theta,$$

and

$$\pi_{\ell} = \frac{\phi + \theta - \sum_{i=1}^{\ell-1} \omega_i \pi_i}{\omega_{\ell}}. \quad (5)$$

Note that the joint prior density of $(\pi_{(\ell)}, \phi)$ in (4) is well defined. We wish to take $\phi = 0$ in (5) to incorporate the restriction, but when $\phi = 0$ the joint density of $\pi_{(\ell)}$ is not well defined.

We assume μ, τ, θ are independent a priori with $p(\mu, \tau, \theta) = p_1(\mu, \tau) p_2(\theta)$, where

$$p_1(\mu, \tau) = \frac{1}{(1+\tau)^2}, \quad 0 < \mu < 1, \tau \geq 0$$

as in (3), and $p_2(\theta)$ is given by

$$\theta \sim \text{Beta}\{\mu_0 \tau_0, (1-\mu_0) \tau_0\}. \quad (6)$$

For the restricted model we consider two scenarios. Letting $\tau_0 \rightarrow \infty$, θ becomes a point mass at μ_0 , and in this case $\theta = \mu_0$ is to be specified by a practitioner; we will call the adjusted model the fixed (FI) model or Model 2. We have a second scenario in which a practitioner specifies μ_0 and τ_0 but not θ ; we will call this adjusted model the informative (IN) model or Model 3. Thus, there are three models, including the unrestricted model. To provide a unified framework, we need all our priors to be proper. The exact value of θ is likely to be unknown in most applications, and this can lead to estimates which are not internally coherent.

It is worth noting that we have considered an additional model to help study the gain in precision of IN relative to FI. For comparison we want to impose a proper but noninformative prior on θ , so that $\theta \sim \text{Uniform}(0, 1)$ is not an unreasonable choice. Letting $\mu_0 = 1/2$, $\tau_0 = 2$, we get $\theta \sim \text{Uniform}(0, 1)$ with this prior, and we will call the adjusted model the uniform (UN) model or Model 4; of

course, we do not need to specify μ_0 and τ_0 . It is worth noting that the prior corresponding to $\tau \rightarrow \infty$ is improper as it corresponds to $\theta \sim \text{Beta}(0, 0)$. We do not consider this model further; however, although UN does not have a constraint, we will consider it briefly throughout.

2.2 Posterior inference

We consider making posterior inference about π_i , $i = 1, \dots, \ell$. Let $\pi = (\pi_1, \dots, \pi_\ell)'$ and $\pi_{(i)} = (\pi_1, \dots, \pi_{i-1}, \pi_{i+1}, \dots, \pi_\ell)'$ [e.g., $\pi_{(\ell)} = (\pi_1, \dots, \pi_{\ell-1})'$ as defined above].

We use Bayes' theorem to find the joint posterior densities of all parameters. First, under the unrestricted model specified by (1), (2) and (3) the joint posterior density of π, μ, τ is

$$g(\pi, \mu, \tau | s) \propto \prod_{i=1}^{\ell} \frac{\pi_i^{s_i + \mu\tau - 1} (1 - \pi_i)^{f_i + (1-\mu)\tau - 1}}{B\{s_i + \mu\tau, f_i + (1-\mu)\tau\}} \\ \times \prod_{i=1}^{\ell} \frac{B\{s_i + \mu\tau, f_i + (1-\mu)\tau\}}{B\{\mu\tau, (1-\mu)\tau\}} \\ \times \frac{1}{(1 + \tau)^2}, \quad (7)$$

$0 < \pi_i < 1, 0 < \mu < 1, \tau > 0, i = 1, \dots, \ell$.

Lemma 1 Under the unrestricted model the joint posterior density, $g(\pi, \mu, \tau | s)$, is proper.

A proof of Lemma 1 is given in Appendix A.

Under the restricted model the joint posterior density of $\pi_{(\ell)}, \mu, \tau, \theta, \phi$ is

$p(\pi_{(\ell)}, \mu, \tau, \theta, \phi | s)$

$$\propto \prod_{i=1}^{\ell-1} \frac{\pi_i^{s_i + \mu\tau - 1} (1 - \pi_i)^{f_i + (1-\mu)\tau - 1}}{B\{s_i + \mu\tau, f_i + (1-\mu)\tau\}} \\ \times \frac{\left[\frac{\phi + \theta - \sum_{i=1}^{\ell-1} \omega_i \pi_i}{\omega_\ell} \right]^{s_\ell + \mu\tau - 1} \left[1 - \frac{\phi + \theta - \sum_{i=1}^{\ell-1} \omega_i \pi_i}{\omega_\ell} \right]^{f_\ell + (1-\mu)\tau - 1}}{B\{s_\ell + \mu\tau, f_\ell + (1-\mu)\tau\}} \\ \times \prod_{i=1}^{\ell} \left[\frac{B\{s_i + \mu\tau, f_i + (1-\mu)\tau\}}{B\{\mu\tau, (1-\mu)\tau\}} \right] \\ \times \theta^{\mu_0 \tau_0 - 1} (1 - \theta)^{(1-\mu_0)\tau_0 - 1} \times \frac{1}{(1 + \tau)^2}, \quad (8)$$

$0 < \pi_i < 1, i = 1, \dots, \ell, 0 < \mu < 1, \tau > 0, \phi + \theta - \omega_\ell \leq \sum_{i=1}^{\ell-1} \omega_i \pi_i \leq \phi + \theta, 0 < \theta < 1$. Note that $\pi_\ell = (\phi + \theta - \sum_{i=1}^{\ell-1} \omega_i \pi_i) / \omega_\ell$.

We get the pertinent joint posterior density by incorporating the constraint ($\phi = 0$) into (8). That is, $p(\pi_{(\ell)}, \mu, \tau, \theta | s, \phi = 0) \propto p(\pi_{(\ell)}, \mu, \tau, \theta, \phi = 0 | s)$, where

$$p(\pi_{(\ell)}, \mu, \tau, \theta | s, \phi = 0) \\ \propto \prod_{i=1}^{\ell-1} \frac{\pi_i^{s_i + \mu\tau - 1} (1 - \pi_i)^{f_i + (1-\mu)\tau - 1}}{B\{s_i + \mu\tau, f_i + (1-\mu)\tau\}} \\ \times \frac{\left[\frac{\theta - \sum_{i=1}^{\ell-1} \omega_i \pi_i}{\omega_\ell} \right]^{s_\ell + \mu\tau - 1} \left[1 - \frac{\theta - \sum_{i=1}^{\ell-1} \omega_i \pi_i}{\omega_\ell} \right]^{f_\ell + (1-\mu)\tau - 1}}{B\{s_\ell + \mu\tau, f_\ell + (1-\mu)\tau\}} \\ \times \prod_{i=1}^{\ell} \left[\frac{B\{s_i + \mu\tau, f_i + (1-\mu)\tau\}}{B\{\mu\tau, (1-\mu)\tau\}} \right] \\ \times \theta^{\mu_0 \tau_0 - 1} (1 - \theta)^{(1-\mu_0)\tau_0 - 1} \times \frac{1}{(1 + \tau)^2}, \quad (9)$$

$0 < \pi_i < 1, i = 1, \dots, \ell, 0 < \mu < 1, \tau > 0, \theta - \omega_\ell \leq \sum_{i=1}^{\ell-1} \omega_i \pi_i \leq \theta, 0 < \theta < 1$. Note again that $\pi_\ell = (\theta - \sum_{i=1}^{\ell-1} \omega_i \pi_i) / \omega_\ell$. It is worth noting that the joint posterior density (9) incorporates the constraint, $\sum_{i=1}^{\ell-1} \omega_i \pi_i = \theta$, exactly because $\pi_\ell = (\theta - \sum_{i=1}^{\ell-1} \omega_i \pi_i) / \omega_\ell, \theta - \omega_\ell \leq \sum_{i=1}^{\ell-1} \omega_i \pi_i \leq \theta$. That is, the joint posterior density is not a function of π_ℓ , and posterior inference about π_i follows from the identity, $\pi_i = (\theta - \sum_{j=1, j \neq i}^{\ell-1} \omega_j \pi_j) / \omega_\ell$. Thus, there is absolutely no difference between θ and $\sum_{i=1}^{\ell-1} \omega_i \pi_i$.

Theorem 1 Under the restricted model the joint posterior density, $p(\pi_{(\ell)}, \mu, \tau, \theta | s, \phi = 0)$, is proper.

A proof of Theorem 1 is given in Appendix A.

We note the difference between the densities for the unrestricted model in (7) and the restricted model in (9). Essentially, the term

$$\left(\frac{\theta - \sum_{i=1}^{\ell-1} \omega_i \pi_i}{\omega_\ell} \right)^{s_\ell + \mu\tau - 1} \times \left(1 - \frac{\theta - \sum_{i=1}^{\ell-1} \omega_i \pi_i}{\omega_\ell} \right)^{f_\ell + (1-\mu)\tau - 1} \\ \times \theta^{\mu_0 \tau_0 - 1} (1 - \theta)^{(1-\mu_0)\tau_0 - 1}$$

in (9) replaces $\pi_\ell^{s_\ell + \mu\tau - 1} (1 - \pi_\ell)^{f_\ell + (1-\mu)\tau - 1}$ in (7). Note that in (9),

$$\pi_\ell = \frac{\theta - \sum_{i=1}^{\ell-1} \omega_i \pi_i}{\omega_\ell}.$$

Let $a_i = s_i + \mu\tau, b_i = f_i + (1-\mu)\tau, i = 1, \dots, \ell$. Also let

$$c_i = \frac{\theta - \sum_{j=1, j \neq i}^{\ell-1} \omega_j \pi_j - \omega_i}{\omega_i}$$

and

$$d_i = \frac{\theta - \sum_{j=1, j \neq i}^{\ell-1} \omega_j \pi_j}{\omega_i}, i = 1, \dots, \ell - 1.$$

Then,

$$p(\pi_i | \pi_{(i)}, \mu, \tau, \theta, s, \phi = 0) \propto \pi_i^{a_i-1} (1 - \pi_i)^{b_i-1} (\pi_i - c_i)^{b_i-1} (d_i - \pi_i)^{a_i-1}, \quad (10)$$

$c_i < \pi_i < d_i, i = 1, \dots, \ell - 1$. Note that this density function consists of two terms $\pi_i^{a_i-1} (1 - \pi_i)^{b_i-1}$ and $(\pi_i - c_i)^{b_i-1} (d_i - \pi_i)^{a_i-1}$; note the interchange between a_i and b_i in the second term. The first term is the conditional posterior density under the unrestricted model, and the second term is a generalized beta density [*i.e.*, a beta(b_i, a_i) distribution in the interval (c_i, d_i)]. Thus, the unrestricted beta density is adjusted by the generalized beta density. In the rest of the paper we denote by $\text{GenBeta}(a, b, c, d)$ the generalized beta random variable with density function,

$$p(x) = (x - c)^{a-1} (d - x)^{b-1} / \{(d - c)^{a+b-1} B(a, b)\},$$

$$c \leq x \leq d, a > 1, b > 1.$$

That is, $(X - c) / (d - c) \sim \text{Beta}(a, b)$ if and only if $X \sim \text{GenBeta}(a, b, c, d)$.

It is worth noting that we have ordered the areas in order of their counts (smallest to largest). This is convenient and advantageous both theoretically and computationally.

In order to explain the gain in precision, we attempt to study (10) further by making two approximations. First, because the restriction under study is rather mild we do not expect c_i to be much different from 0 and d_i to be much different from 1. Under this assumption, we can approximate (10) by

$$p_a(\pi_i | \pi_{(i)}, \mu, \tau, \theta, s, \phi = 0)$$

$$\propto (\pi_i - c_i)^{a_i-1} (d_i - \pi_i)^{b_i-1} (\pi_i - c_i)^{b_i-1} (d_i - \pi_i)^{a_i-1},$$

$$c_i < \pi_i < d_i.$$

Then, incorporating the normalization constant into $p_a(\pi_i | \pi_{(i)}, \mu, \tau, \theta, s, \phi = 0)$, we have

$$\begin{aligned} p_a(\pi_i | \pi_{(i)}, \mu, \tau, \theta, s, \phi = 0) &= \frac{(\pi_i - c_i)^{a_i-1} (d_i - \pi_i)^{b_i-1} (\pi_i - c_i)^{b_i-1} (d_i - \pi_i)^{a_i-1}}{\int_{c_i}^{d_i} (\pi_i - c_i)^{a_i-1} (d_i - \pi_i)^{b_i-1} (\pi_i - c_i)^{b_i-1} (d_i - \pi_i)^{a_i-1} d\pi_i} \\ &= \frac{(\pi_i - c_i)^{a_i-1} (d_i - \pi_i)^{b_i-1}}{(d_i - c_i)^{a_i+b_i-1} B(a_i, b_i)} \\ &\times \frac{(\pi_i - c_i)^{b_i-1} (d_i - \pi_i)^{a_i-1}}{E[(\pi_i - c_i)^{b_i-1} (d_i - \pi_i)^{a_i-1}]}, c_i < \pi_i < d_i, \end{aligned} \quad (11)$$

where the expectation is taken over the generalized Beta distribution $\pi_i \sim \text{GenBeta}(a_i, b_i, c_i, d_i), i = 1, \dots, \ell - 1$. But under this latter density, $(\pi_i - c_i)^{b_i-1} (d_i - \pi_i)^{a_i-1}$ is an unbiased estimator of $E[(\pi_i - c_i)^{b_i-1} (d_i - \pi_i)^{a_i-1}]$. In addition, by construction a_i and b_i are relatively large and therefore $(\pi_i - c_i)^{b_i-1} (d_i - \pi_i)^{a_i-1}$ and its variance are expected to be small. Then, our second approximation is

$$(\pi_i - c_i)^{b_i-1} (d_i - \pi_i)^{a_i-1} \approx E[(\pi_i - c_i)^{b_i-1} (d_i - \pi_i)^{a_i-1}]. \quad (12)$$

Therefore, combining (11) and (12), our final approximation of (10) is

$$\pi_i | \pi_{(i)}, \mu, \tau, \theta, s, \phi = 0 \sim \text{GenBeta}(a_i, b_i, c_i, d_i). \quad (13)$$

It follows from (13) that

$$E_r(\pi_i | \pi_{(i)}, \mu, \tau, \theta, s, \phi = 0) \approx c_i + (d_i - c_i) E_u(\pi_i | \mu, \tau, s)$$

and

$$\begin{aligned} \text{Var}_r(\pi_i | \pi_{(i)}, \mu, \tau, \theta, s, \phi = 0) &\approx (d_i - c_i)^2 \text{Var}_u(\pi_i | \mu, \tau, s), \end{aligned} \quad (14)$$

where u refers to the unrestricted model and r restricted model. Note that when $c_i = 0$ and $d_i = 1$, we get $E_r(\pi_i | \cdot) = E_u(\pi_i | \cdot)$ and $\text{Var}_r(\pi_i | \cdot) = \text{Var}_u(\pi_i | \cdot)$. Generally though the estimates of π_i will be a bit different from one scenario to the other. It is also interesting that $\text{Var}_r(\pi_i | \cdot) \leq \text{Var}_u(\pi_i | \cdot)$ at least approximately. Thus, the restriction $\sum_{i=1}^{\ell} \omega_i \pi_i = \theta$ will reduce variability, when the π_i are estimated. This is true because the $\pi_i, i = 1, \dots, \ell$, belong to an $\ell - 1$ dimensional simplex in the ℓ dimensional hypercube while for the unrestricted model $\pi_i, i = 1, \dots, \ell$, belong to the ℓ dimensional hypercube. We expect the largest gain in precision when θ is completely specified, followed by the case when μ_0 is specified and $\tau_0 \gg 2$, and the least gain in precision when $\theta \sim \text{Uniform}(0, 1)$.

2.3 Computation

We show how to draw samples from the unrestricted and restricted models. For the unrestricted model we are able to draw random samples from (7) without using Markov chain Monte Carlo methods. However, for the restricted model we use the griddy Gibbs sampler (Ritter and Tanner 1992) to draw samples from (9).

2.3.1 Unrestricted model

We collapse over the π_i , draw samples from $p(\mu, \tau | s)$ using random draws from a bivariate grid, and finally obtain samples from the Rao-Blackwellized densities $\pi_i | \mu, \tau, s$.

Then,

$$\pi_i | \mu, \tau, s \sim \text{Beta}\{s_i + \mu\tau, f_i + (1 - \mu)\tau\}, i = 1, \dots, \ell, \quad (15)$$

and integrating out π , we get

$$p(\mu, \tau | s) \propto \prod_{i=1}^{\ell} \frac{B\{s_i + \mu\tau, f_i + (1 - \mu)\tau\}}{B\{\mu\tau, (1 - \mu)\tau\}} \times \frac{1}{(1 + \tau)^2},$$

$0 < \mu < 1, \tau > 0$. Letting $\delta = \tau / \tau + 1$, we have

$$p(\mu, \delta | s) \propto \left[\prod_{i=1}^{\ell} \frac{B\{s_i + \mu\tau, f_i + (1 - \mu)\tau\}}{B\{\mu\tau, (1 - \mu)\tau\}} \right]_{\tau = \frac{\delta}{1 - \delta}}, \quad 0 < \mu, \delta < 1.$$

First we draw $\mu, \delta | s$ using a bivariate grid on $(0, 1)^2$ to obtain a sample of $M \approx 10,000$ values of $(\mu^{(h)}, \delta^{(h)})$, $h = 1, \dots, M$, $\tau^{(h)} = \delta^{(h)} / (1 - \delta^{(h)})$. Then we perform a data augmentation in (15) to obtain $\pi^{(h)}$, $h = 1, 2, \dots, M$, using a composition method. That is, we simply draw $\pi_i \sim \text{Beta}\{s_i + \mu^{(h)}\tau^{(h)}, f_i + (1 - \mu^{(h)})\tau^{(h)}\}$, $i = 1, \dots, \ell$, $h = 1, \dots, M$.

To perform the bivariate grid method for sampling from the posterior density of (μ, δ) , we divide the interval $(0, 1)$ into 100 sub-intervals; so there are 10,000 little squares in the original unit square. We obtain the heights of the posterior density (without the normalization constant) at the center of each of the 10,000 squares. Because these little squares have the same area, the heights of the bivariate density are proportional to the posterior probabilities that (μ, δ) fall in each of these squares. Thus, we have constructed a joint posterior probability mass function of (μ, δ) on very fine grids. It is easy to draw a sample from the discrete bivariate probability mass function by using the cumulative distribution method. This is actually a random draw of one of the 10,000 squares with probabilities proportional to the heights of the little squares. Then within the selected square we choose a point at random by drawing two uniform random variables (*i.e.*, uniform random jittering). Indeed, this is a very accurate random draw from the joint posterior density of (μ, δ) . We draw $M = 10,000$ samples from this approximation for posterior inference in a standard Monte Carlo procedure with independent samples, not a Markov chain. Because of the random jittering the numbers are different with probability one.

2.3.2 Restricted model

We show how to draw samples from the restricted model using the Gibbs sampler. The joint conditional posterior density of $\pi_1, \dots, \pi_{\ell-1}$ is

$$\begin{aligned} p(\pi_1, \dots, \pi_{\ell-1} | \mu, \tau, \theta, s, \phi = 0) \\ \propto \prod_{i=1}^{\ell-1} \left\{ \pi_i^{s_i + \mu\tau - 1} (1 - \pi_i)^{f_i + (1 - \mu)\tau - 1} \right\} \\ \times \left(\theta - \sum_{i=1}^{\ell-1} \omega_i \pi_i \right)^{s_{\ell} + \mu\tau - 1} \left\{ \sum_{i=1}^{\ell-1} \omega_i \pi_i - \theta + \omega_{\ell} \right\}^{f_{\ell} + (1 - \mu)\tau - 1} \end{aligned} \quad (16)$$

where

$$\theta < \omega_{\ell}, \theta - \omega_{\ell} < \sum_{i=1}^{\ell-1} \omega_i \pi_i < \theta, \pi_{\ell} = \frac{\theta - \sum_{i=1}^{\ell-1} \omega_i \pi_i}{\omega_{\ell}}.$$

Thus, we would obtain samples of $\pi_1, \dots, \pi_{\ell-1}$ and we set

$$\pi_{\ell} = \frac{\left(\theta - \sum_{i=1}^{\ell-1} \omega_i \pi_i \right)}{\omega_{\ell}}$$

to complete the vector π_1, \dots, π_{ℓ} . That is, the constraint is obtained exactly. The conditional posterior density of θ is

$$\begin{aligned} p(\theta | \pi_{(\ell)}, \mu, \tau, s, \phi = 0) \\ \propto \left\{ \theta - \sum_{i=1}^{\ell-1} \omega_i \pi_i \right\}^{s_{\ell} + \mu\tau - 1} \left\{ \omega_{\ell} + \sum_{i=1}^{\ell-1} \omega_i \pi_i - \theta \right\}^{f_{\ell} + (1 - \mu)\tau - 1} \\ \times \theta^{\mu_0\tau_0 - 1} (1 - \theta)^{(1 - \mu_0)\tau_0 - 1}, \end{aligned} \quad (17)$$

where

$$\sum_{i=1}^{\ell-1} \omega_i \pi_i < \theta < \omega_{\ell} + \sum_{i=1}^{\ell-1} \omega_i \pi_i.$$

The joint conditional posterior density of μ and τ is

$$\begin{aligned} p(\mu, \tau | \pi_{(\ell)}, \theta, s, \phi = 0) \\ \propto \frac{q^{\mu\tau} r^{(1 - \mu)\tau}}{[B(\mu\tau, (1 - \mu)\tau)]^{\ell}} \times \frac{1}{(1 + \tau)^2}, \end{aligned} \quad (18)$$

$$0 < \mu < 1, \tau > 0, q = \prod_{i=1}^{\ell} \pi_i, r = \prod_{i=1}^{\ell} (1 - \pi_i).$$

To perform the Gibbs sampler, we need to draw samples from (16), (17) and (18), each in turn, until convergence. We draw μ, τ from $p(\mu, \tau | \pi_{(\ell)}, \theta, s)$ in a manner similar to drawing from $p(\mu, \tau | \pi_{(\ell)})$ in the unrestricted model. It is more difficult to draw sample from (16) and (17). However, we use essentially the same method to draw samples from the conditional posterior density of $\pi_i, i = 1, \dots, \ell - 1$, obtained from (16) and θ from (17) which are both proportional to the product of two density functions, one is a truncated beta density and the other a generalized beta density. We next develop some theory to draw a sample from such a density. For this purpose, we state and prove Lemma 2 and Theorem 2.

The density function of interest is

$$f(x) = Af_1(x)f_2(x), 0 \leq c < x < d \leq 1, \quad (19)$$

where

$$f_1(x) = \frac{x^{g-1}(1-x)^{h-1}}{\int_c^d x^{g-1}(1-x)^{h-1} dx}, \quad c < x < d, g, h > 0, \quad (20)$$

$$f_2(x) = (x-c)^{a-1}(d-x)^{b-1} / \{(d-c)^{a+b-1}B(a, b)\},$$

$$c < x < d, a, b > 1, \quad (21)$$

and, of course,

$$A = 1 / \int_c^d f_1(x)f_2(x) dx. \quad (22)$$

It is worth noting that we are not assuming $g, h > 1$. If this was the case, then $f_1(x)$ and $f_2(x)$ will be both logconcave, thereby making $f(x)$ logconcave, and in this case one can draw a sample from $f(x)$ using the adaptive rejection sampler (ARS, Gilks and Wild 1992). We are providing a specialized algorithm to draw a sample from $f(x)$ which is not logconcave. Even if $f_1(x)$ was logconcave (*i.e.*, $g, h > 1$) this specialized algorithm will still be better than the ARS because the ARS is a general purpose algorithm; see Robert and Casella (1999, page 59). Our algorithm requires less computation and does not need logconcavity; even if there is logconcavity the ARS can perform poorly in the tails of the density function.

Lemma 2 Consider the density functions $f_1(x)$ and $f_2(x)$ with $a, b > 1$.

(a) Then

$$\sup_{c < x < d} f_2(x) = \frac{\delta^{a-1}(1-\delta)^{b-1}}{(d-c)B(a, b)}, \delta = (a-1)/(a+b-2).$$

(b) For any $g > 0, h > 0$ there exist two constants H_1 and H_2 such that

$$0 < H_1 \leq A^{-1} \leq H_2 < \infty.$$

A proof of Lemma 2 is given in Appendix A.

Theorem 2 Let $F_{g,h}(\cdot)$ be the cdf of Beta(g, h) random variable and $F_{g,h}^{-1}(\cdot)$ be its inverse. Let

$$U, V \sim \text{Uniform}(0, 1),$$

and let

$$X = F_{g,h}^{-1}\{UF_{g,h}(d) + (1-U)F_{g,h}(c)\}.$$

If for two real numbers $a, b > 1$,

$$V \leq \frac{1}{(d-c)^{a+b-2}} \left(\frac{X-c}{\delta} \right)^{a-1} \left(\frac{d-X}{1-\delta} \right)^{b-1},$$

where $\delta = (a-1)/(a+b-2)$, then X has the density $f(x) = Af_1(x)f_2(x)$.

A proof of Theorem 2 is given in Appendix A.

Theorem 1 gives us the following algorithm for drawing samples from $f(\pi) \propto \pi^{g-1}(1-\pi)^{h-1}(\pi-c)^{a-1}(d-\pi)^{b-1}$, $c < \pi < d, g, h > 0, a, b > 1$.

Algorithm

(a) Draw $U \sim \text{Uniform}(0, 1)$ and set

$$\pi = F_{g,h}^{-1}\{UF_{g,h}(d) + (1-U)F_{g,h}(c)\}.$$

(b) Draw $V \sim \text{Uniform}(0, 1)$. If

$$V \leq \frac{1}{(d-c)^{a+b-2}} \left(\frac{\pi-c}{\delta} \right)^{a-1} \left(\frac{d-\pi}{1-\delta} \right)^{b-1},$$

accept π , otherwise go to (a).

Because the binomial sample sizes are arranged in increasing order, in any application it will be true that $a, b > 1$ and $g, h > 0$ (possibly greater than 1 as well). Thus, the algorithm will work. Indeed, in all our examples (one presented here) and simulation exercises the algorithm runs very quickly.

Now, we show how to draw $\pi_i, i = 1, \dots, \ell$, and θ . For π_i ,

$$p(\pi_i | \pi_{(i,\ell)}, \theta, \mu, \tau, s, \phi = 0)$$

$$\propto \pi_i^{a_i-1}(1-\pi_i)^{b_i-1}(\pi_i - c_i)^{b_\ell-1}(d_i - \pi_i)^{a_\ell-1}, c_i < \pi_i < d_i,$$

where $\pi_{(i,\ell)}$ is the vector containing the elements of π except for π_i and π_ℓ , and $a_i = s_i + \mu\tau, b_i = f_i + (1-\mu)\tau, i = 1, \dots, \ell$,

$$c_i = \left(\theta - \sum_{j=1, j \neq i}^{\ell-1} \omega_j \pi_j - \omega_\ell \right) / \omega_i,$$

$$d_i = \left(\theta - \sum_{j=1, j \neq i}^{\ell-1} \omega_j \pi_j \right) / \omega_i, i = 1, \dots, \ell-1.$$

Apply the theorem to $p(\pi_i | \pi_{(i)}, \theta, \mu, \tau, s), a_\ell > 1, b_\ell > 1, i = 1, \dots, \ell-1$.

For θ , we have

$$p(\theta | \pi, \mu, \tau, s, \phi = 0)$$

$$\propto \theta^{\mu_0 \tau_0 - 1} (1 - \theta)^{(1 - \mu_0) \tau_0 - 1} (\theta - \tilde{c})^{a_\ell - 1} (\tilde{d} - \theta)^{b_\ell - 1}, \tilde{c} < \theta < \tilde{d},$$

where

$$\tilde{c} = \sum_{i=1}^{\ell-1} \omega_i \pi_i, \quad \tilde{d} = \omega_\ell + \sum_{i=1}^{\ell-1} \omega_i \pi_i.$$

Again, apply the theorem, $a_\ell > 1$, $b_\ell > 1$.

When θ is fully specified (*i.e.*, θ is not random), we do not have to draw θ . However, when $\theta \sim \text{Uniform}(0, 1)$ a priori ($\mu_0 = 1/2$, $\tau_0 = 2$), we have a simplification. In this case,

$$\theta | (\pi_{(\ell)}, \mu, \tau, s, \phi = 0) \sim \text{GenBeta}(a_\ell, b_\ell, \tilde{c}, \tilde{d})$$

and $\theta = \tilde{c} + (\tilde{d} - \tilde{c})X$, where $X \sim \text{Beta}(a_\ell, b_\ell)$, has the required density.

For both the unrestricted and restricted models we use 10,000 iterates to make posterior inference about the binomial probabilities, π_i . Under the unrestricted model these are simply random draws and no monitoring is required. For the restricted model, running the griddy Gibbs sampler, we drew 11,000 iterates, used 1,000 as a “burn in” (a conservative number because convergence occur much earlier as evident in the trace plots) and we found negligible correlations among the iterates. Thus, we used 10,000 iterates to make inference about the binomial probabilities. For both the unrestricted and the three restricted models it takes only a few seconds on our 2×833 MHz alpha computer.

3. Numerical studies

In Section 3.1 we describe an illustrative example to show the main features of the restriction. In Section 3.2 we describe a simulation study to show frequentist properties of the Bayes estimators, and we show deeper insight into the differences among the four scenarios. Note again that when we performed the computations, it is convenient to order the domain sizes so that the largest domain comes last.

3.1 Illustrative example

We have used data in the third National Health and Nutrition Examination (NHANES) Survey to illustrate our method. We have studied body mass index for teenagers, and we have data on the sample obtained. The domains (small areas) are formed by crossing ethnicity (white, black, Mexican) and sex (male, female). We have separated out the teenagers with respect to whether they were in middle

school or high school at the time of the survey. Thus, there are 12 small domains. The data are presented in the first four columns of Table 1 by domain. Note that domains MWM, MBF, MWF and HBF are relatively sparse with 4, 2, 5, 5 obese teenagers respectively; for the twelve domains the sample consists of 959 with 130 obese teenagers (*i.e.*, the overall proportion of obese individuals is 0.136 approximately). In column 4 of Table 1 we have also presented the direct estimates by domains, and these estimates range from 0.069 to 0.228. The estimates for the smallest domains will be unreliable. Moreover, when the beta-binomial models are used, these estimates will regress to the overall sample mean of 0.136, creating a possible bias. Our method is expected to increase precision beyond the unrestricted model because the restricted model uses more information about the weighted sum. Clearly, predictors based on either the restricted model or the beta-binomial model are biased if the specified model is wrong.

We have taken $\mu_0 = 0.136$, the overall sample proportion, and $\tau_0 = 959$, the total sample size. Less optimistic choices can be used. For example, $\tau_0 = 100$, say; but this choice makes very little difference. However, it is worth noting that using the observed data to specify the prior distribution can artificially decrease the posterior variance. Typically a survey practitioner will have an appropriate specification from a prior survey or a census. One cannot specify values for μ_0 and τ_0 which are completely out of line and will create huge biases. Here τ_0 is a prior sample size and μ_0 is a prior mean of θ . This method permits a sensible value for θ ; we are essentially adding a degree of uncertainty about knowledge of the linear combination. Thus, these specifications are not unreasonable.

We have applied our method as described for the four scenarios. In the other columns of Table 1 we study the estimates of the small area probabilities. We present the posterior mean (PM), posterior standard deviation (PSD),

$$\text{RMSE} = \sqrt{(\hat{\pi} - \text{PM})^2 + \text{PSD}^2},$$

where $\hat{\pi}$ is the direct estimate, and the 95% highest posteriori density (HPD) interval (Int). As is expected, the PSDs are roughly in the increasing order: Model 2, Model 3, Model 4 and Model 1; in some cases the differences are important. The PMs for Models 1, 2 and 3 are mostly similar, but for Model 4 the PMs are mostly smaller than the other three models. There is much improvement of Models 2 and 3 over Model 1 at least in terms of precision. This gain becomes less important for Model 4, the model with the greatest uncertainty about θ .

Table 1

Comparison of the four models using posterior mean (PM), posterior standard deviation, root mean square error (RMSE), and 95% credible HPD intervals (Int) of π_i by domain (D) for the NHANES data

D	s	n	$\hat{\pi}$	PM	PSD	RMSE	Int	PM	PSD	RMSE	Int
				Model 1				Model 2			
1	4	47	0.085	0.114	0.033	0.044	(0.051, 0.179)	0.111	0.032	0.041	(0.049, 0.170)
2	2	29	0.069	0.112	0.037	0.057	(0.042, 0.183)	0.111	0.036	0.055	(0.041, 0.178)
3	10	44	0.227	0.175	0.044	0.068	(0.100, 0.264)	0.177	0.041	0.065	(0.108, 0.260)
4	5	62	0.081	0.107	0.030	0.040	(0.047, 0.159)	0.107	0.027	0.038	(0.054, 0.160)
5	10	74	0.135	0.134	0.030	0.030	(0.077, 0.194)	0.134	0.028	0.028	(0.080, 0.190)
6	12	69	0.174	0.158	0.036	0.039	(0.089, 0.227)	0.155	0.031	0.036	(0.095, 0.214)
7	8	79	0.101	0.116	0.028	0.031	(0.065, 0.173)	0.115	0.027	0.030	(0.065, 0.166)
8	5	62	0.081	0.107	0.030	0.040	(0.052, 0.169)	0.105	0.029	0.038	(0.042, 0.153)
9	28	123	0.228	0.196	0.036	0.048	(0.129, 0.262)	0.196	0.032	0.045	(0.131, 0.253)
10	10	111	0.090	0.106	0.026	0.030	(0.059, 0.155)	0.105	0.024	0.028	(0.061, 0.150)
11	16	122	0.131	0.132	0.026	0.026	(0.083, 0.183)	0.130	0.023	0.023	(0.090, 0.179)
12	20	137	0.146	0.144	0.026	0.026	(0.094, 0.194)	0.141	0.022	0.023	(0.100, 0.184)
				Model 3				Model 4			
1	4	47	0.085	0.111	0.033	0.042	(0.044, 0.169)	0.109	0.032	0.040	(0.050, 0.172)
2	2	29	0.069	0.111	0.037	0.056	(0.039, 0.179)	0.108	0.036	0.053	(0.037, 0.173)
3	10	44	0.227	0.175	0.043	0.068	(0.093, 0.260)	0.170	0.044	0.072	(0.091, 0.255)
4	5	62	0.081	0.106	0.029	0.038	(0.050, 0.160)	0.103	0.030	0.038	(0.048, 0.164)
5	10	74	0.135	0.134	0.029	0.029	(0.077, 0.189)	0.129	0.030	0.030	(0.067, 0.184)
6	12	79	0.174	0.156	0.034	0.038	(0.090, 0.217)	0.151	0.036	0.043	(0.087, 0.222)
7	8	69	0.101	0.118	0.028	0.033	(0.062, 0.171)	0.111	0.028	0.029	(0.061, 0.167)
8	5	62	0.081	0.107	0.030	0.040	(0.051, 0.165)	0.102	0.030	0.036	(0.050, 0.159)
9	28	123	0.228	0.195	0.034	0.047	(0.138, 0.265)	0.189	0.035	0.052	(0.123, 0.255)
10	10	111	0.090	0.107	0.024	0.029	(0.062, 0.156)	0.104	0.025	0.029	(0.051, 0.149)
11	16	122	0.131	0.132	0.024	0.024	(0.086, 0.179)	0.126	0.025	0.025	(0.083, 0.179)
12	20	137	0.146	0.143	0.024	0.024	(0.095, 0.191)	0.137	0.025	0.027	(0.091, 0.189)

Note: The four models are: Model 1 - no restriction; Model 2 - fixed θ ; Model 3 - informative prior for θ ; Model 4 - uniform prior for θ . Domains are formed by crossing school (middle school - M, high school - H), race (white - W, black - B, mexican american - M) and sex (male - M, female - F). Thus, the domains are: 1-MWM, 2-MBF, 3-MMM, 4-MWF, 5-MBM, 6-MMF, 7-HWM, 8-HBF, 9-HMM, 10-HWF, 11-HBM, 12-HMF (e.g., the first domain consists of middle school white boys). n is the number of teenagers and s the number of obese teenagers in each domain. Data are taken from the 35 largest counties in the US. An estimate of the overall probability is $130 / 959 \approx 0.136$, and for the first domain $\hat{p} = 4 / 47 = 0.085$; the numerical standard errors are all smaller than 0.001; $RMSE = \sqrt{(\hat{\pi} - PM)^2 + PSD^2}$.

We also study very briefly the nuisance parameter θ . We note that the weighted average of the direct estimators of the small areas is 0.136 (more accurately 0.1355599). When θ is held fixed at 0.1355599, the weighted average of the posterior means is 0.136. When θ has the informative prior, the weighted average of the posterior means is 0.136, and for θ the PM is 0.136, the PSD is 0.008, and a 95% HPD interval for θ is (0.122, 0.152). When θ has the uniform prior, the weighted average of the posterior means is 0.132, and for θ the PM is 0.131, the PSD is 0.011, and a 95% HPD interval for θ is (0.110, 0.151). This shows the deficiencies of the uniform prior which we use only for comparison. It is worth noting that μ_1, \dots, μ_{t-1} and θ are computed first. Then μ_t is obtained by subtraction. This is done at each iterate of the Gibbs sampler. Then, the posterior summaries for $\sum_{i=1}^t \omega_i \pi_i$ and θ are computed. So there will be very minor discrepancies which are due to rounding.

Finally, we have selected the four smallest domains to compare the posterior densities of the probabilities. We have used the Parzen-Rosenblatt kernel density estimator to estimate the posterior densities; see Silverman (1986) for details. Figure 1 compares the estimated posterior densities for the four models. It is interesting that as the domain sizes increase, the four models get closer together. Also, for all cases the tails of the distributions in each panel are very similar; the differences in these distributions though lie in the modal intervals (i.e., interval containing the mode), and their heights. As expected, the posterior density corresponding to the unrestricted model is the shortest, simply because it has more variability. Model 4 has posterior density shifted to the left and is slightly bimodal for the smallest domain. Thus, inference about the modes of these distributions will be different. But inference involving the tails will not be so different; except for Model 4, 95% credible intervals will be similar.

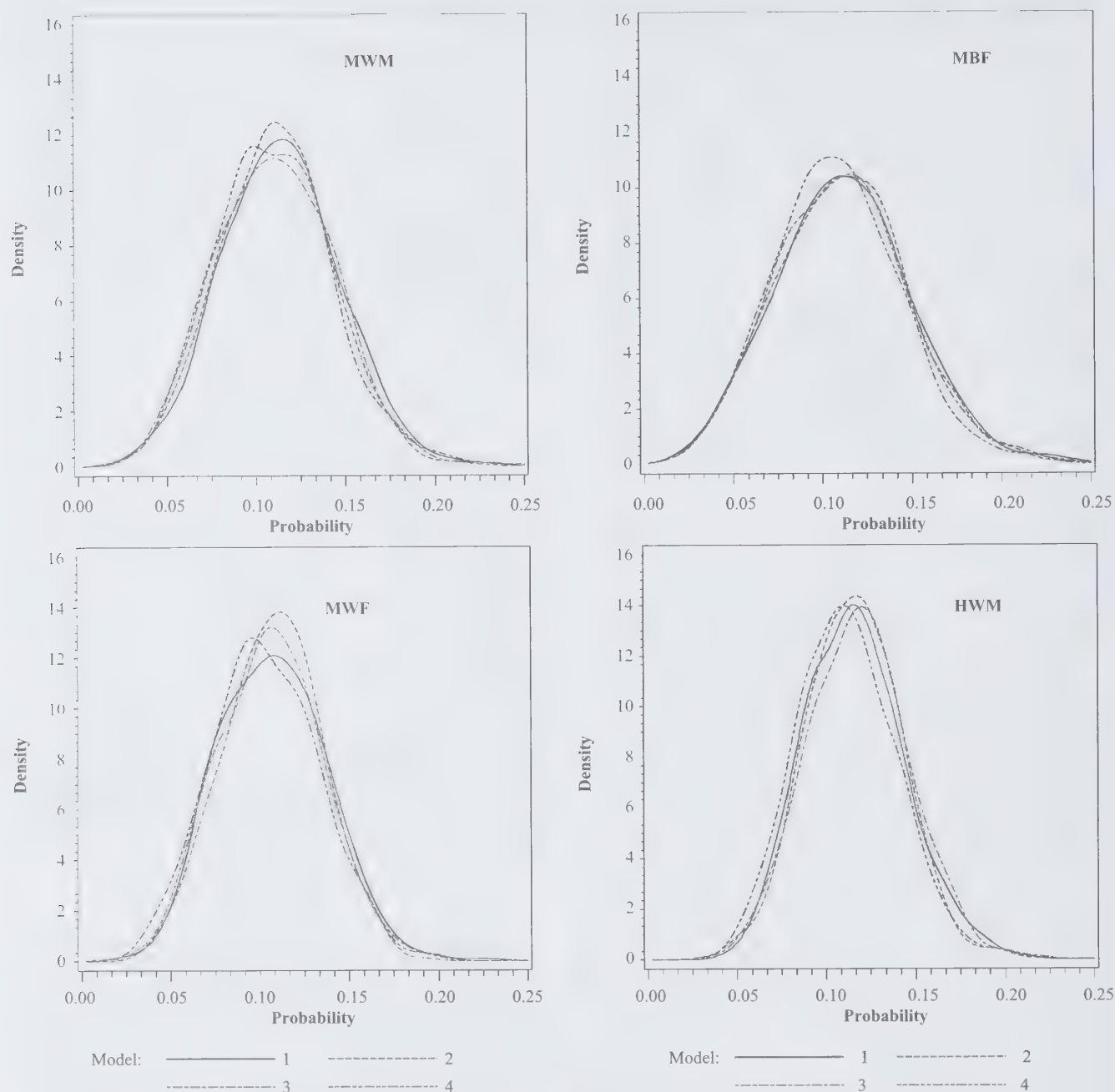


Figure 1 Plots of the estimated posterior densities of π_1 , π_2 , π_4 , and π_7 for the four models and NHANES data

3.2 Simulation study

We use a simulation study to assess the statistical properties of our method. We want to see if the gain in precision persists and to see how the estimators of the probabilities are shifted. We also study the frequentist properties of the estimators of the probabilities. In the description of the simulation it is convenient to use the abbreviated names of the models which are UR (Model 1, no restriction), FI (Model 2, fixed θ), IN (Model 3,

informative prior for θ) and UN (Model 4, uniform prior for θ).

We set $\theta_0 = 0.15$, $\mu_0 = \theta_0$ and $\tau_0 = 100$. We have selected three values of $\ell = 12, 24, 36$, 12 being the number of areas in the NHANES data. We drew the sample sizes from a uniform density in $(25, 150)$, again to reflect the NHANES data. First, we generated

$$\pi_i \sim \text{iid Beta} \{ \mu_0 \tau_0, (1 - \mu_0) \tau_0 \}, i = 1, \dots, \ell.$$

To do this latter task, we drew sets of $\ell \pi_i$ until $\theta_0 - w_\ell \leq \sum_{i=1}^{\ell-1} \omega_i \pi_i \leq \theta_0$; set $\pi_\ell = (\theta_0 - \sum_{i=1}^{\ell-1} \omega_i \pi_i) / w_\ell$. Then, we generated

$$s_i^{\text{ind}} \sim \text{Binomial}(n_i, \pi_i).$$

We have generated 1,000 data sets in this manner for each of $\ell = 12, 24, 36$. Then, we fit the four models (one unrestricted and three restricted models). The process is very fast (*i.e.*, for samples sizes of 12, 24, 30 there were respectively 22, 90, 153 rejects in the 1,000 samples). We fit each data set using random samples for the unrestricted model and the gridy Gibbs sampler for the restricted models. We fit the 1,000 data sets in a couple of hours on our 2×833 MHz alpha computer.

For these 1,000 simulations we study PM, the coverage (C), the bias (B), PSD, RMSE and width (W) of the 95% credible intervals. For each domain we compute the bias $\text{PM} - \pi$, then we average these values over all domains and simulation runs, and this quantity we now call B . Associated with B we also computed AB , the average of $|\text{PM} - \pi|$. Similarly, we have computed

$$\text{RMSE} = \sqrt{(\text{PM} - \pi)^2 + \text{PSD}^2}$$

for each domain and each simulation run and we average these over all domains and simulation runs. Note that the true probabilities, π_i , are known by design. We obtain the coverage (C) by computing the proportion of all intervals containing the true value of π_i over all domains and simulation runs. We also obtain the average of the widths of the 95% credible intervals. Numerical standard errors are obtained for all quantities.

Table 2

Simulation: Comparison of the four models using coverage (C), bias and average absolute bias (B and AB), posterior standard deviation (PSD), root mean squared error (RMSE) and width of the 95% credible intervals (W) of π_i

ℓ	Model	C	B	AB	PSD	RMSE	W
12	UR	0.960 _{0.0018}	-0.002 _{0.0003}	0.0231 _{0.00016}	0.033 _{0.0001}	0.043 _{0.0001}	0.125 _{0.0003}
	FI	0.961 _{0.0018}	-0.000 _{0.0003}	0.0219 _{0.00020}	0.031 _{0.0001}	0.040 _{0.0001}	0.118 _{0.0003}
	IN	0.946 _{0.0021}	0.005 _{0.0003}	0.0275 _{0.00066}	0.032 _{0.0001}	0.043 _{0.0001}	0.122 _{0.0002}
	UN	0.956 _{0.0019}	-0.000 _{0.0003}	0.0261 _{0.00019}	0.032 _{0.0001}	0.042 _{0.0001}	0.122 _{0.0003}
24	UR	0.957 _{0.0013}	-0.001 _{0.0002}	0.0229 _{0.00012}	0.031 _{0.0000}	0.041 _{0.0001}	0.119 _{0.0002}
	FI	0.957 _{0.0013}	-0.000 _{0.0002}	0.0224 _{0.00013}	0.030 _{0.0000}	0.040 _{0.0001}	0.116 _{0.0002}
	IN	0.943 _{0.0015}	0.006 _{0.0002}	0.0252 _{0.00058}	0.030 _{0.0000}	0.041 _{0.0001}	0.116 _{0.0001}
	UN	0.952 _{0.0014}	-0.000 _{0.0002}	0.0236 _{0.00012}	0.031 _{0.0002}	0.041 _{0.0002}	0.118 _{0.0005}
36	UR	0.960 _{0.0010}	-0.001 _{0.0001}	0.0224 _{0.00009}	0.030 _{0.0000}	0.040 _{0.0001}	0.117 _{0.0001}
	FI	0.961 _{0.0010}	-0.000 _{0.0001}	0.0218 _{0.00009}	0.030 _{0.0000}	0.039 _{0.0001}	0.115 _{0.0001}
	IN	0.948 _{0.0012}	0.005 _{0.0002}	0.0224 _{0.00009}	0.030 _{0.0000}	0.040 _{0.0001}	0.114 _{0.0001}
	UN	0.957 _{0.0011}	-0.000 _{0.0001}	0.0228 _{0.00010}	0.030 _{0.0000}	0.040 _{0.0001}	0.116 _{0.0001}

Note: The four models are: Model 1 - no restriction (UR); Model 2 - fixed θ (FI); Model 3 - informative prior for θ (IN); Model 4 - uniform prior for θ (UN). $\text{RMSE} = \sqrt{(\pi - \text{PM})^2 + \text{PSD}^2}$. The notation a_b means a is an estimate and b is the standard error.

In Table 2 we study the estimates of the small area probabilities. It is convenient to use the shorter names of the four models for our discussion. For IN the PMs are close to the nominal value of 0.15, but for UN the PMs are smaller than the nominal value particularly for UN at $\ell = 12$. We observe that the coverage for all the models UR, FI and UN are always larger than the nominal value of 95%, but for model IN these coverages are smaller than the nominal value of 95%. A similar difference exists for the bias; while the bias is small for all models, models UR, FI (the specified value of θ is 0.15) and UN have negative biases but IN has positive bias. Except for $\ell = 36$ IN has the largest AB. The PSDs are mostly similar and the RMSEs share the same features; there are some differences at $\ell = 12$. The four models get similar as ℓ increases; when ℓ is large there appears to be no need for our method. However, again the gain in precision appears to be in the increasing order FI, IN, UN and UR.

In most applications the exact value of θ is unknown. Therefore, the PSDs of the π_i , under the situation where θ is assumed known, are likely to underestimate the true PSDs. So we study the deviations of the PSDs of IN and UN from those of FI, and we compute the ratios, $R_1 = \text{PSD}_{\text{IN}} / \text{PSD}_{\text{FI}}$ and $R_2 = \text{PSD}_{\text{UN}} / \text{PSD}_{\text{FI}}$. In Table 3 we present the five-number summaries of these ratios by sample size. Most of the ratios are around 1 (*i.e.*, inter-quartile range) with some tendency for them to be larger than 1. (Note that the maxima at $\ell =$ and $\ell = 24$ are outliers possibly due to bad simulated samples.) Thus, overall the PSDs under IN and UN are not much larger under FI.

Table 3
Simulation: A study of the posterior standard deviation (PSD) of the π_i using five number summaries of the ratios, R_1 and R_2 , by sample size

ℓ	Ratio	Min	Q_1	Med	Q_3	Max
12	R_1	0.673	0.972	1.032	1.091	5.329
	R_2	0.022	0.984	1.034	1.086	85.370
24	R_1	0.019	0.965	1.005	1.047	16.017
	R_2	0.024	0.979	1.014	1.049	486.960
36	R_1	0.690	0.962	0.998	1.034	1.236
	R_2	0.837	0.979	1.011	1.044	1.243

Note: $R_1 = \text{PSD}_{\text{IN}} / \text{PSD}_{\text{FI}}$ and $R_2 = \text{PSD}_{\text{UN}} / \text{PSD}_{\text{FI}}$. The five summaries are minimum (min), first quartile (Q_1), median (med), third quartile (Q_3) and maximum (max).

In Table 4 we study the estimate of θ for the two pertinent models IN and UN. For both models the coverage probabilities are smaller than the nominal value, and the coverage for UN is smaller than the interval for IN. Bias is small for both models, positive for IN and negative for UN.

Table 4
Simulation: Comparison of the informative (IN) and the uniform (UN) models using posterior mean (PM), coverage (C), bias and average absolute bias (B and AB), posterior standard deviation (PSD), root mean squared error (RMSE) and width of the 95% credible intervals (W) of π_i

ℓ	Model	PM	C	B	AB	PSD	RMSE	W
12	IN	0.149 _{0.0012}	0.853 _{0.0112}	0.000 _{0.0003}	0.00152 _{0.00081}	0.008 _{0.0000}	0.012 _{0.0002}	0.030 _{0.0001}
	UN	0.138 _{0.0005}	0.881 _{0.0102}	-0.012 _{0.0004}	0.00038 _{0.00003}	0.011 _{0.0001}	0.016 _{0.0002}	0.042 _{0.0002}
24	IN	0.153 _{0.0015}	0.833 _{0.0118}	0.003 _{0.0015}	0.00212 _{0.00103}	0.007 _{0.0006}	0.012 _{0.0015}	0.024 _{0.0015}
	UN	0.145 _{0.0029}	0.842 _{0.0115}	-0.005 _{0.0003}	0.00012 _{0.00006}	0.008 _{0.0001}	0.012 _{0.0002}	0.030 _{0.0002}
36	IN	0.150 _{0.0002}	0.828 _{0.0119}	0.000 _{0.0002}	0.00004 _{0.00000}	0.004 _{0.0000}	0.007 _{0.0001}	0.017 _{0.0001}
	UN	0.145 _{0.0003}	0.794 _{0.0128}	-0.005 _{0.0002}	0.00009 _{0.00000}	0.006 _{0.0000}	0.010 _{0.0001}	0.024 _{0.0001}

Note: The two models considered are: Model 3 – informative prior for θ and model 4 – uniform prior for θ . $\text{RMSE} = (\theta_0 - \text{PM})^2 + \text{PSD}^2$. The notation a_b means a is an estimate and b is the standard error.

Table 5
Simulation: Comparison of the four models using posterior standard deviation and root mean square error (RMSE) of π_i by domain (D)

D	Unrestricted		Fixed		Informative		Uniform	
	PSD	RMSE	PSD	RMSE	PSD	RMSE	PSD	RMSE
1	0.048 _{0.0003}	0.057 _{0.0004}	0.046 _{0.0003}	0.054 _{0.0004}	0.045 _{0.0002}	0.056 _{0.0005}	0.047 _{0.0004}	0.056 _{0.0005}
2	0.046 _{0.0003}	0.055 _{0.0004}	0.044 _{0.0003}	0.053 _{0.0004}	0.044 _{0.0002}	0.054 _{0.0005}	0.045 _{0.0004}	0.054 _{0.0005}
3	0.044 _{0.0002}	0.053 _{0.0004}	0.042 _{0.0002}	0.050 _{0.0004}	0.042 _{0.0002}	0.052 _{0.0005}	0.043 _{0.0003}	0.051 _{0.0004}
4	0.042 _{0.0002}	0.050 _{0.0004}	0.040 _{0.0002}	0.047 _{0.0004}	0.040 _{0.0002}	0.050 _{0.0004}	0.041 _{0.0002}	0.049 _{0.0004}
5	0.041 _{0.0002}	0.049 _{0.0004}	0.038 _{0.0002}	0.046 _{0.0004}	0.039 _{0.0002}	0.048 _{0.0004}	0.039 _{0.0003}	0.048 _{0.0005}
6	0.040 _{0.0002}	0.048 _{0.0004}	0.037 _{0.0002}	0.045 _{0.0004}	0.037 _{0.0002}	0.048 _{0.0004}	0.038 _{0.0003}	0.047 _{0.0005}
7	0.038 _{0.0002}	0.046 _{0.0004}	0.035 _{0.0002}	0.043 _{0.0003}	0.036 _{0.0002}	0.046 _{0.0004}	0.037 _{0.0003}	0.045 _{0.0004}
8	0.037 _{0.0002}	0.045 _{0.0003}	0.034 _{0.0002}	0.041 _{0.0003}	0.036 _{0.0002}	0.046 _{0.0004}	0.036 _{0.0003}	0.044 _{0.0004}
9	0.036 _{0.0002}	0.044 _{0.0003}	0.033 _{0.0002}	0.040 _{0.0004}	0.034 _{0.0001}	0.044 _{0.0004}	0.035 _{0.0003}	0.042 _{0.0004}
10	0.035 _{0.0002}	0.043 _{0.0003}	0.032 _{0.0002}	0.039 _{0.0003}	0.034 _{0.0001}	0.044 _{0.0004}	0.034 _{0.0003}	0.042 _{0.0004}
11	0.034 _{0.0001}	0.042 _{0.0003}	0.031 _{0.0002}	0.038 _{0.0003}	0.033 _{0.0001}	0.042 _{0.0004}	0.033 _{0.0003}	0.041 _{0.0004}
12	0.035 _{0.0002}	0.047 _{0.0005}	0.031 _{0.0002}	0.042 _{0.0004}	0.034 _{0.0003}	0.047 _{0.0006}	0.034 _{0.0007}	0.046 _{0.0008}

Note: The four models are: Model 1 – no restriction; Model 2 – fixed θ ; Model 3 – informative prior for θ ; Model 4 – uniform prior for θ . $\text{RMSE} = \sqrt{(\pi_i - \text{PM})^2 + \text{PSD}^2}$. The notation a_b means a is an estimate and b is the standard error. Here 12 domains are used and the original simulated sample sizes are divided by 2.

Except for $\ell = 36$ IN has by far the larger AB. The PSDs and RMSEs are generally smaller for IN, and the widths of the 95% credible intervals are significantly smaller for IN. It appears that it is difficult to estimate θ under UN, but IN appears to be somewhat better.

In Table 5 we present more detailed result (*i.e.*, by domain) for the case when the number of domains is 12. To show further gains in precision, we have reduced the sample size to half as much [*i.e.*, we drew the sample sizes uniformly in the interval (12, 75)]. We present the posterior standard deviation and the posterior root mean square error, averaged over the simulation runs. Again the standard errors are presented. We note that all the probability contents (not presented) are at least the nominal value of 95%. The numerical standard errors are small in all cases. The PSDs and RMSEs are in the right order. Note that because the sample sizes are arranged in order from smallest to largest, there is a decrease in the PSDs and RMSEs as the domain numbers go up.

We study the posterior density of π_1 for $\ell = 12$, and we compare the four models. Again we use the Parzen-Rosenblatt density estimator. In Figure 2 we present the estimated posterior densities (Parzen-Rosenblatt) averaged over the 1,000 runs for $\ell = 12$. We obtain the same results as for the BMI data. Again the tails are similar. FI is the tallest density and UN is the shortest. UN is slightly shifted to the left of IN. In Figure 3 we present a systematic sample of 10 densities from the 1,000 simulation runs by model. We can see large variation among the 10 estimated posterior densities. Again we can see that FI is tallest; UR, FI and UN show similar variation with IN slightly taller. Thus, it is important to take the average for comparison as in Figure 2.

4. Concluding remarks

We have extended the beta-binomial model of small area estimation to accommodate a prior specification of a weighted average of the area probabilities. We have used the Bayesian approach which is particularly attractive for problems with awkward likelihood functions as in our application with the constraint of the weighted average of the beta-binomial model. We viewed the constraint as prior knowledge which can be precise or less informative. The griddy Gibbs sampler is used to fit the models, thereby avoiding the more sophisticated Metropolis-Hastings sampler. We have developed a theory which permits sampling from a density function which is proportional to the product of a truncated beta-binomial density and a generalized beta density. We have found that overall our complete algorithm forming the griddy Gibbs sampler runs efficiently and fast.

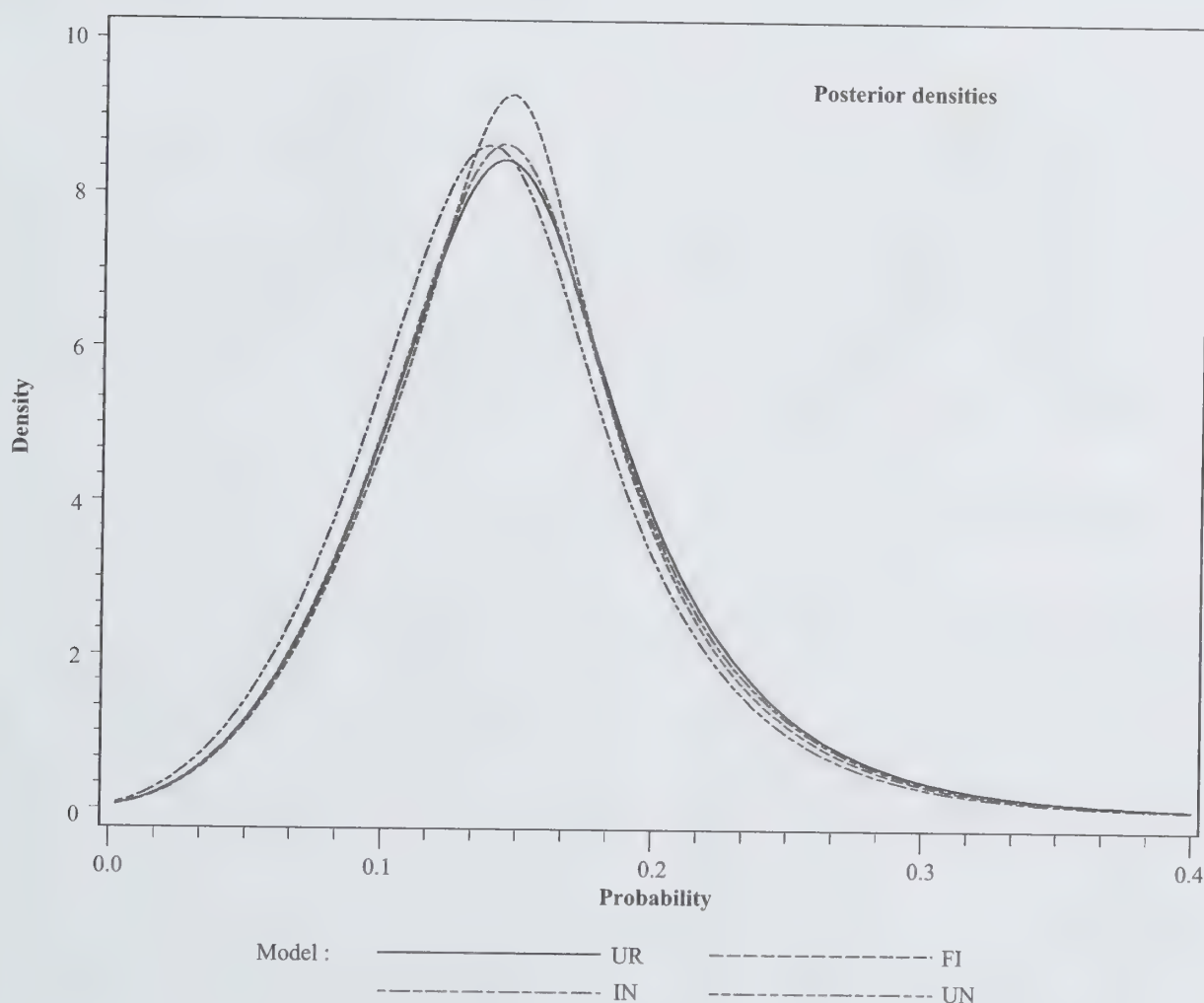


Figure 2 Plots of the estimated posterior densities of π_1 by model when there are 12 domains

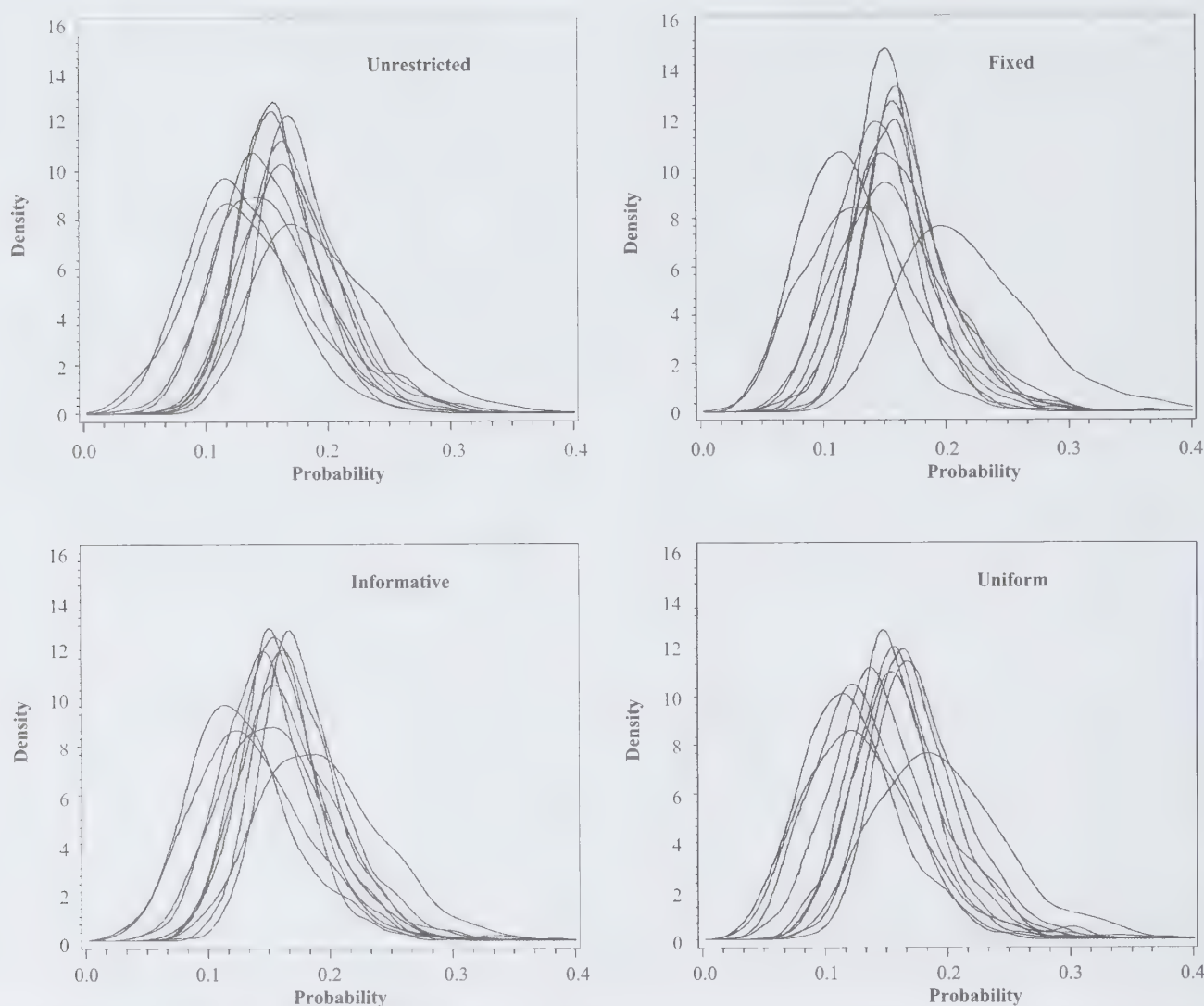


Figure 3 Plots of the estimated posterior densities of π_1 for a systematic sample of size 10 from the 1,000 runs by model when there are 12 domains

We have shown that there could be gains in precision when extra information is incorporated into the beta-binomial model. We have considered three scenarios in which a survey practitioner (a) can not specify any constraint (standard beta-binomial model for small areas), (b) can specify a constraint and the parameter completely, and (c) can specify a constraint and information which can be used to construct a prior distribution for the parameter. Our example on obesity of children in the National Health and Nutrition Examination Survey and simulation study showed that the gain in precision beyond (a) is in an order with (b) larger than (c). As the exact algebraic arguments are difficult, we obtained an analytical approximation which shows that indeed there could be gain in precision of (b) over (a). For comparison we have considered a fourth scenario in which θ has vague information, and as expected, it turned to be rather uninteresting and inefficient.

It is straight forward to make Bayesian predictive inference about the finite population mean of each small area. Let $P_i = T_i / N_i$ denote the finite population proportion for the i^{th} area, where $T_i = \sum_{j=1}^{N_i} y_{ij}$, y_{ij} are the binary responses, and N_i , the number of individuals in the i^{th} area, is assumed known. Now $T_i = t_i^{(s)} + t_i^{(ns)}$, where $t_i^{(s)}$ and $t_i^{(ns)}$ are respectively the sample total and the nonsample total. Now under any of the models $t_i^{(ns)} | \pi_i \sim \text{Binomial}(n_i, \pi_i)$ and $p(t_i^{(ns)} | y_s) = \int p(t_i^{(ns)} | \pi_i) p(\pi_i | y_s) d\pi_i$, where $y_s = (y_1, \dots, y_\ell)'$. Thus, it is easy to obtain the empirical posterior density of P_i using a sampling-based method. Nandram and Sedransk (1993) obtained some analytical features of P_i when τ is known, but not with the constraint; see also Nandram (1998).

We mention a generalization of our restricted beta-binomial hierarchical Bayesian model to the Dirichlet-multinomial model (e.g., Nandram 1998). Let y_i be c -vector of

cell counts (*i.e.*, number of people possessing one of c traits), and let n_i denote the sample sizes within the i^{th} area, $i = 1, \dots, \ell$. We assume

$$y_i | \pi_i \sim \text{Multinomial}(n_i, \pi_i), \pi_i | \mu, \tau, \theta \stackrel{\text{iid}}{\sim} \text{Dirichlet}(\mu\tau)$$

with $\sum_{i=1}^{\ell} w_i \pi_i = \theta$. Finally $\theta \sim \text{Dirichlet}(\mu_0 \tau_0)$, where μ_0 and τ_0 are to be specified, and independently $p(\mu, \tau) = (k-1)!/(1+\tau)^2$, $0 < \mu_k < 1$, $k = 1, \dots, c$, $\sum_{k=1}^c \mu_k = 1$. With k constraints this problem is much more complex, but we plan to work on it. Other extensions to nonignorable non-response (Nandram and Choi 2002) and two-way categorical tables are possible.

Acknowledgements

The authors are grateful to the Associate Editor and the two referees who helped enormously to improve the quality of the presentation.

Appendix A

Proofs of lemmas 1, 2 and theorems 1, 2

Proof of lemma 1

This is a special case of a general result. Using the multiplication rule and because the prior is proper, it is clear that the joint density of π, μ, τ, s “integrates” to one. Therefore, the joint posterior density of π, μ, τ given s is proper.

Proof of theorem 1

Let $\mathcal{T} = \{(\pi, \mu, \tau, \theta): 0 < \pi_i < 1, i = 1, \dots, \ell, 0 < \mu < 1, \tau > 0, \theta - \omega_{\ell} \leq \sum_{i=1}^{\ell-1} \omega_i \pi_i \leq \theta, 0 < \theta < 1, \pi_{\ell} = (\theta - \sum_{i=1}^{\ell-1} \omega_i \pi_i) / \omega_{\ell}\}$ and $\mathcal{T}^* = \{(\pi, \mu, \tau): 0 < \pi_i < 1, i = 1, \dots, \ell, 0 < \mu < 1, \tau > 0\}$; note that $\mathcal{T} \subset \mathcal{T}^*$.

Let $\tilde{g}(\pi, \mu, \tau | s)$ denote the right-hand side of the unrestricted posterior density in (7) and $\tilde{p}(\pi_{(\ell)}, \mu, \tau, \theta | s, \phi = 0)$ denote the right-hand side of the restricted posterior density in (9). Noting that $\pi_{\ell} = (\theta - \sum_{i=1}^{\ell-1} \omega_i \pi_i) / \omega_{\ell}$, we observe that

$$\tilde{p}(\pi_{(\ell)}, \mu, \tau, \theta | s, \phi = 0) =$$

$$\tilde{g}(\pi, \mu, \tau | s) \times \theta^{\mu_0 \tau_0 - 1} (1 - \theta)^{(1 - \mu_0) \tau_0 - 1}, (\pi, \mu, \tau, \theta) \in \mathcal{T}.$$

Because $\theta^{\mu_0 \tau_0 - 1} (1 - \theta)^{(1 - \mu_0) \tau_0 - 1}$ is proportional to the density function of beta random variable, we have

$$\int_{\mathcal{T}} \tilde{p}(\pi_{(\ell)}, \mu, \tau, \theta | s, \phi = 0) d\pi d\mu d\tau d\theta =$$

$$A \int_{\mathcal{T}} \tilde{g}(\pi, \mu, \tau | s) d\pi d\mu d\tau \leq A \int_{\mathcal{T}^*} \tilde{g}(\pi, \mu, \tau | s) d\pi d\mu d\tau,$$

where $A = B\{\mu_0 \tau_0, (1 - \mu_0) \tau_0\}$ is the beta function. By lemma 1, $\int_{\mathcal{T}^*} \tilde{g}(\pi, \mu, \tau | s) d\pi d\mu d\tau < \infty$. Thus, $p(\pi_{(\ell)}, \mu, \tau, \theta | s, \phi = 0)$ is proper.

Proof of Lemma 2 (a)

This can be proved in two ways. The second derivative of $\log\{f_2(x)\}$ is negative in (c, d) , and so the first derivative, when set to zero, provides a unique mode which is $\delta d + (1 - \delta)c$. Alternatively, because $(X - c)/(d - c) \sim \text{Beta}(a, b)$ with $a, b > 1$, there is a unique mode for $(X - c)/(d - c)$, and this translates to $\delta d + (1 - \delta)c$; note that $\delta d + (1 - \delta)c$ is a point in (c, d) . Thus, substituting $\delta d + (1 - \delta)c$ into $f_2(x)$, we have

$$\sup_{c < x < d} f_2(x) = \delta^{a-1} (1 - \delta)^{b-1} / (d - c) B(a, b).$$

Proof of Lemma 2 (b)

Because $a, b > 1$, $x \geq x - c$ and $1 - x \geq d - x$, it is true that

$$A^{-1} \geq D^{-1} \int_c^d (x - c)^{a+g-2} (d - x)^{b+h-2} dx,$$

where $D = (d - c)^{a+b-1} B(a, b) B(g, h) \{F_{g,h}(d) - F_{g,h}(c)\}$ and $F_{g,h}(x)$ is the cdf of a standard beta random variable in $(0, 1)$. Note that because $c < d$ (strictly) and $F_{g,h}(x)$ is monotone increasing in $(0, 1)$, $F_{g,h}(d) - F_{g,h}(c) > 0$ (strictly). By comparison with the generalized beta density [*i.e.*, $\text{Beta}(a + g - 1, b + h - 1, c, d)$], the integral is $(d - c)^{a+b+g+h-3} B(a + g - 1, b + h - 1)$. Thus,

$$A^{-1} \geq \frac{(d - c)^{g+h-2} B(a + g - 1, b + h - 1)}{B(a, b) B(g, h) \{F_{g,h}(d) - F_{g,h}(c)\}} = H_1 > 0.$$

Also, we have

$$A^{-1} \leq \int_c^d f_1(x) \sup_{c < x < d} f_2(x) dx.$$

Then by Lemma 2 (a),

$$\begin{aligned} A^{-1} &\leq \frac{\delta^{a-1} (1 - \delta)^{b-1}}{(d - c) B(a, b)} \int_c^d f_1(x) dx = \frac{\delta^{a-1} (1 - \delta)^{b-1}}{(d - c) B(a, b)} \\ &= H_2 < \infty. \end{aligned}$$

Proof of theorem 2

To show the claim, we calculate the cdf $F_X(\cdot)$ of the random variable X defined in the Theorem. We have

$$\begin{aligned} F_X(x) &= P(X \leq x) \\ &= P[F_{g,h}^{-1}\{UF_{g,h}(d) + (1 - U)F_{g,h}(c)\} \leq x] \\ &= P[UF_{g,h}(d) + (1 - U)F_{g,h}(c) \leq F_{g,h}(x)] \\ &= P[U\{F_{g,h}(d) - F_{g,h}(c)\} \leq F_{g,h}(x) - F_{g,h}(c)] \\ &= P\left[U \leq \frac{F_{g,h}(x) - F_{g,h}(c)}{F_{g,h}(d) - F_{g,h}(c)}\right]. \end{aligned}$$

Now, since $U \sim \text{Uniform}(0, 1)$, from the above expression for $F_X(\cdot)$, we have $F_X(x) = 1$ if $x \geq d$ and $F_X(x) = 0$ if $x \leq c$. When $c \leq x \leq d$, we have

$$F_X(x) = \frac{F_{g,h}(x) - F_{g,h}(c)}{F_{g,h}(d) - F_{g,h}(c)}.$$

This shows that X has the truncated beta density $f_1(x)$ in (20).

Now, looking to use the accept-reject algorithm, consider

$$\frac{f(x)}{f_1(x)} = Af_2(x).$$

By Lemma 2, we have

$$\sup_{c < \pi < d} \left\{ \frac{f(x)}{f_1(x)} \right\} = A \sup_{c < \pi < d} f_2(x) = A \frac{\delta^{a-1}(1-\delta)^{b-1}}{(d-c)B(a, b)} < \infty.$$

Thus, by the accept-reject algorithm, if

$$V \leq \frac{1}{(d-c)^{a+b-2}} \left(\frac{X-c}{\delta} \right)^{a-1} \left(\frac{d-X}{1-\delta} \right)^{b-1},$$

then X has the density $f(x)$ in (19).

References

- Cochran, W.G. (1977). *Sampling Techniques*, third edition. New York: John Wiley & Sons, Inc.
- Gilks, W.R., and Wild, P. (1992). Adaptive rejection sampling for gibbs sampling. *Journal of the Royal Statistical Society, Series C*, 41, 337-348.
- Gelman, A. (2006). Prior distribution for variance parameters in hierarchical models. *Bayesian Analysis*, 1, 515-533.
- Ghosh, M., Natarajan, K., Stroud, T.W.F. and Carlin, B.P. (1998). Generalized linear models for small-area estimation. *Journal of the American Statistical Association*, 93, 273-282.
- Hillmer, S.C., and Trabelsi, A. (1987). Benchmarking of economic time series. *Journal of the American Statistical Association*, 82, 1064-1071.
- Lazar, R., Meeden, G. and Nelson, D. (2008). A noninformative Bayesian approach to finite population sampling using auxiliary variables. *Survey Methodology*, 34, 51-64.
- Nandram, B. (1998). A Bayesian analysis of the three-stage hierarchical multinomial model. *Journal of Statistical Computation and Simulation*, 61, 97-126.
- Nandram, B., and Choi, J.W. (2002). Hierarchical Bayesian nonresponse models for binary data from small areas with uncertainty about ignorability. *Journal of the American Statistical Association*, 97, 381-388.
- Nandram, B., and Choi, J.W. (2002). A Bayesian analysis of a proportion under non-ignorable non-response. *Statistics in Medicine*, 21, 9, 1189-1212.
- Nandram, B., and Sedransk, J. (1993). Bayesian predictive inference for a finite population proportion: Two-stage cluster sampling. *Journal of the Royal Statistical Society, Series B*, 55, 399-408.
- Nandram, B., Toto, M.C.S. and Choi, J.W. (2011). A Bayesian benchmarking of the Scott-Smith model for small areas. *Journal of Statistical Computation and Simulation* (in press, preprint).
- Rao, J.N.K. (2003). *Small Area Estimation*. New York: John Wiley & Sons, Inc.
- Ritter, C., and Tanner, M.A. (1992). The gibbs sampler and the griddy gibbs sampler. *Journal of the American Statistical Association*, 87, 861-868.
- Robert, C.P., and Casella, G. (1999). *Monte Carlo Statistical Methods*. New York: Springer-Verlag.
- Silvapulle, M.J., and Sen, P.K. (2006). *Constrained Statistical Inference: Inequality, Order and Shape Restrictions*. New York: John Wiley & Sons, Inc.
- Silverman, B.W. (1986). *Density Estimation*. London: Chapman and Hall.

On bias-robust mean squared error estimation for pseudo-linear small area estimators

Ray Chambers, Hukum Chandra and Nikos Tzavidis¹

Abstract

We propose a method of mean squared error (MSE) estimation for estimators of finite population domain means that can be expressed in pseudo-linear form, *i.e.*, as weighted sums of sample values. In particular, it can be used for estimating the MSE of the empirical best linear unbiased predictor, the model-based direct estimator and the M-quantile predictor. The proposed method represents an extension of the ideas in Royall and Cumberland (1978) and leads to MSE estimators that are simpler to implement, and potentially more bias-robust, than those suggested in the small area literature. However, it should be noted that the MSE estimators defined using this method can also exhibit large variability when the area-specific sample sizes are very small. We illustrate the performance of the method through extensive model-based and design-based simulation, with the latter based on two realistic survey data sets containing small area information.

Key Words: Best linear unbiased prediction; M-quantile model; Model-based direct estimation; Random effects model; Small area estimation.

1. Introduction

Linear models, and linear predictors based on these models, are widely used in survey-based inference. However, such models run the risk of misspecification, particularly with regard to second order and higher moments. Bias-robust methods for estimating the mean squared error (MSE) of linear predictors of finite population quantities, *i.e.*, methods that remain approximately unbiased under failure of assumptions about second order and higher moments, have been developed. Valliant, Dorfman and Royall (2000, Chapter 5) discuss bias-robust MSE estimation for such predictors when a population is assumed to follow a linear model.

In this paper we address a subsidiary problem, which is that of bias-robust MSE estimation for estimators of finite population domain means that can be expressed in pseudo-linear form, *i.e.*, as weighted sums, but where the weights can depend on the sample values of the variable of interest. An important application, and one that motivates our approach, is small area inference. Consequently from now on we use 'area' to refer to a domain of interest. Our approach represents an extension of the ideas in Royall and Cumberland (1978) and appears to lead to simpler to implement MSE estimators than those that have been suggested in the small area literature.

The structure of the paper is as follows. In section 2 we discuss MSE estimation under an area-specific linear model. That is, we focus on estimation of the conditional MSE. We then show how our approach can be used for estimating the MSE of three different small area linear predictors when

they are expressed in pseudo-linear form, (a) the empirical best linear unbiased predictor or EBLUP (Henderson 1953); (b) the model-based direct estimator (MBDE) of Chandra and Chambers (2009); and (c) the M-quantile predictor (Chambers and Tzavidis 2006). In section 3 we present results from a series of simulation studies that illustrate the model-based and the design-based properties of our approach to MSE estimation. Finally, in section 4 we summarize our main findings. Throughout, we use either i or h to index the D small areas of interest, and either j or k to index the distinct population units in these areas.

2. Bias-robust MSE estimation for pseudo-linear estimators

2.1 MSE estimation under an area-specific linear model

We consider the situation where we have a finite population of size N from which a sample of size n is drawn. We assume that this population consists of D non-overlapping domains, each one of which contains sampled units, with small realised sample sizes in each of the sampled domains. As noted earlier, and following standard practice, we refer to these domains as areas from now on. We assume also that there is a known number N_i of population units in area i , with n_i of these sampled. The total number of units in the population is $N = \sum_{i=1}^D N_i$, with corresponding total sample size $n = \sum_{i=1}^D n_i$. In what follows, we use s to denote the collection of units in sample, with s_i the subset drawn from area i , and use expressions like $j \in i$ and $j \in s$ to refer to the units making up area i and sample s respectively.

1. Ray Chambers, Centre for Statistical and Survey Methodology, University of Wollongong, Wollongong, NSW, 2522, Australia. E-mail: ray@uow.edu.au; Hukum Chandra, Indian Agricultural Statistics Research Institute, Library Avenue, New Delhi-110012, India. E-mail: hchandra@iasri.res.in; Nikos Tzavidis, Social Statistics and Southampton Statistical Sciences Research Institute, University of Southampton, Southampton, SO17 1BJ, UK. E-mail: n.tzavidis@soton.ac.uk.

Linear models are often used to motivate estimators for population means. However, when estimates are required for the corresponding area means, it is usually not realistic to assume that a linear model that applies to the population as a whole also applies within each area. We therefore adopt a conditional approach, and consider MSE estimation for estimators of area means when different linear models apply within different areas. In particular, we focus on estimators that can be expressed as weighted sums of the sample values, referring to them as 'linear' in what follows to indicate that they have a linear structure.

To start, let y_j denote the value of Y for unit j of the population and suppose that this unit is in area i . We also assume an area-specific linear model for y_j of the form

$$y_j = \mathbf{x}_j^T \boldsymbol{\beta}_i + e_j. \quad (1)$$

Here \mathbf{x}_j is a $p \times 1$ vector of unit level auxiliary variables for unit j , $\boldsymbol{\beta}_i$ is a $p \times 1$ vector of area-specific regression coefficients and e_j is a unit level random effect with mean zero and variance σ_j^2 that is uncorrelated between different population units. We do not make any assumptions about σ_j^2 at this point. Note that throughout this paper we assume that the sampling method used is non-informative for the population values of Y given the corresponding values of the auxiliary variables and knowledge of the area affiliations of the population units. As a consequence, (1) applies at both sample and population level.

Let \mathbf{y}_s denote the column vector of sample values of y_j and let $\mathbf{w}_{is} = \{w_{ij}; j \in s\}$ denote the column vector of fixed weights such that $\hat{m}_i = \mathbf{w}_{is}^T \mathbf{y}_s = \sum_{j \in s} w_{ij} y_j$ is a linear estimator of $m_i = N_i^{-1} \sum_{j \in i} y_j$. By 'fixed' here we mean that these weights do not depend on the sample values of Y . Moreover, we assume $w_{ij} = O(n_i^{-1})$ for $j \in s_i$, $w_{ij} = o(n_i^{-1})$ for $j \notin s_i$, and $\sum_{j \in s} w_{ij} = 1$. Here s_i denotes the n_i sample units from area i . The bias of \hat{m}_i under (1) is then

$$E(\hat{m}_i - m_i) = \left(\sum_{h=1}^D \sum_{j \in s_h} w_{ij} \mathbf{x}_j^T \boldsymbol{\beta}_h \right) - \bar{\mathbf{x}}_i^T \boldsymbol{\beta}_i, \quad (2)$$

where $\bar{\mathbf{x}}_i$ denotes the vector of average values of the auxiliary variables in area i . Similarly, the prediction variance of \hat{m}_i under (1) is

$$\text{Var}(\hat{m}_i - m_i) = N_i^{-2} \left\{ \sum_{h=1}^D \sum_{j \in s_h} a_{ij}^2 \sigma_j^2 + \sum_{j \in r_i} \sigma_j^2 \right\}, \quad (3)$$

where r_i denotes the non-sampled units in area i and $a_{ij} = N_i w_{ij} - I(j \in i)$. We use $I(A)$ to denote the indicator function for event A , so $I(j \in i)$ takes the value 1 if population unit j is from area i and is zero otherwise. Note that since a_{ij} is $O(N_i n_i^{-1})$ for $j \in s_i$, the first term within the braces in (3) is the leading term of this prediction variance if N_i is large compared to n_i .

Let $j \in h$. We consider the important special case where $\mu_j = E(y_j | \mathbf{x}_j) = \mathbf{x}_j^T \boldsymbol{\beta}_h$ is estimated by $\hat{\mu}_j = \mathbf{x}_j^T \hat{\boldsymbol{\beta}}_h = \sum_{k \in s} \phi_{kj} y_k$, with the ϕ_{kj} corresponding to suitable weights. Then

$$y_j - \hat{\mu}_j = (1 - \phi_{jj}) y_j - \sum_{k \in s(-j)} \phi_{kj} y_k$$

and so

$$\text{Var}(y_j - \hat{\mu}_j) = \sigma_j^2 \left\{ (1 - \phi_{jj})^2 + \sum_{k \in s(-j)} \phi_{kj}^2 (\sigma_k^2 / \sigma_j^2) \right\} \quad (4)$$

under (1). Here $s(-j)$ denotes the sample s with unit j excluded. If in addition $\hat{\mu}_j$ is unbiased for μ_j under (1), i.e.,

$$E(y_j - \hat{\mu}_j) = 0, \quad (5)$$

we can then adopt the approach of Royall and Cumberland (1978) and estimate (3) by

$$\hat{V}(\hat{m}_i) = N_i^{-2} \left\{ \sum_{h=1}^D \sum_{j \in s_h} a_{ij}^2 \hat{\lambda}_j^{-1} (y_j - \hat{\mu}_j)^2 + \sum_{j \in r_i} \hat{\sigma}_j^2 \right\}, \quad (6)$$

where $\hat{\lambda}_j = (1 - \phi_{jj})^2 + \sum_{k \in s(-j)} \phi_{kj}^2$ and $\hat{\sigma}_j^2 = \hat{\sigma}_k^2 / \hat{\sigma}_j^2$. Usually, the estimates $\hat{\sigma}_j^2$ of the residual variances in (6) are derived under a 'working model' refinement to (1). In the situation of most concern to us, where the sample sizes within the different areas are too small to reliably estimate area-specific variability, a pooling assumption can be made, i.e., $\sigma_j^2 = \sigma^2$, in which case we put

$$\hat{\sigma}_j^2 = \hat{\sigma}^2 = n^{-1} \sum_{j \in s} \left\{ (1 - \phi_{jj})^2 + \sum_{k \in s(-j)} \phi_{kj}^2 \right\}^{-1} (y_j - \hat{\mu}_j)^2.$$

In this case (6) becomes

$$\hat{V}(\hat{m}_i) = N_i^{-2} \sum_{j \in s} \left\{ a_{ij}^2 + (N_i - n_i) n^{-1} \right\} \hat{\lambda}_j^{-1} (y_j - \hat{\mu}_j)^2, \quad (7)$$

where now $\hat{\lambda}_j = (1 - \phi_{jj})^2 + \sum_{k \in s(-j)} \phi_{kj}^2$. Since any assumptions regarding σ_j^2 in the working model extension of (1) only affect second order terms in (3), the estimator (7) is bias-robust, i.e., it remains approximately unbiased under misspecification of the second order moments of this working model.

A corresponding estimator of the MSE of \hat{m}_i under (1) follows directly. This is

$$\hat{M}(\hat{m}_i) = \hat{V}(\hat{m}_i) + \hat{B}^2(\hat{m}_i), \quad (8)$$

where

$$\hat{B}(\hat{m}_i) = \sum_{h=1}^D \sum_{j \in s_h} w_{ij} \hat{\mu}_j - N_i^{-1} \sum_{j \in i} \hat{\mu}_j \quad (9)$$

is the obvious unbiased estimator of (2).

Use of the square of the unbiased estimator (9) of the bias of \hat{m}_i in the conditional MSE estimator (8) can be criticised because this term is not itself unbiased for the squared bias term in MSE. This can be corrected by replacing (9) by

$$\hat{M}(\hat{m}_i) = \hat{V}(\hat{m}_i) + \hat{B}^2(\hat{m}_i) - \hat{V}\{\hat{B}(\hat{m}_i)\}, \quad (10)$$

where $\hat{V}\{\hat{B}(\hat{m}_i)\}$ is a suitable estimator of the variance of (9). However, we do not recommend use of (10). To see this, let $\bar{\beta} = D^{-1} \sum_{h=1}^D \hat{\beta}_h$ and put $\mathbf{d}_h = \hat{\beta}_h - \bar{\beta}$, where $\hat{\beta}_h$ is the estimator of β_h implied by the weights ϕ_{kj} . Furthermore, put $w_{hi} = \sum_{j \in s_h} w_{ij}$ and $\bar{\mathbf{x}}_{whi} = w_{hi}^{-1} \sum_{j \in s_h} w_{ij} \mathbf{x}_j$, so $\bar{\mathbf{x}}_{wi} = \sum_{h=1}^D \sum_{j \in s_h} w_{ij} \mathbf{x}_j = \sum_{h=1}^D w_{hi} \bar{\mathbf{x}}_{whi}$ is the estimate of $\bar{\mathbf{x}}_i$ based on the weights w_{ij} . Finally, let $\delta_{hi} = \bar{\mathbf{x}}_h^T \mathbf{d}_h - \bar{\mathbf{x}}_i^T \mathbf{d}_i$ and put $\delta_i = \sum_{h=1}^D w_{hi} \delta_{hi}$. Then (9) can be written

$$\begin{aligned} \hat{B}(\hat{m}_i) &= (\bar{\mathbf{x}}_{wi} - \bar{\mathbf{x}}_i)^T \bar{\beta} \\ &+ \left(\sum_{h=1}^D w_{hi} \bar{\mathbf{x}}_{whi}^T \mathbf{d}_h - \bar{\mathbf{x}}_i^T \mathbf{d}_i \right) \\ &= (\bar{\mathbf{x}}_{wi} - \bar{\mathbf{x}}_i)^T \bar{\beta} \\ &+ \left(\sum_{h=1}^D w_{hi} (\bar{\mathbf{x}}_{whi} - \bar{\mathbf{x}}_h)^T \mathbf{d}_h + \sum_{h=1}^D w_{hi} \bar{\mathbf{x}}_h^T \mathbf{d}_h - \bar{\mathbf{x}}_i^T \mathbf{d}_i \right) \\ &= (\bar{\mathbf{x}}_{wi} - \bar{\mathbf{x}}_i)^T \bar{\beta} \\ &+ \sum_{h=1}^D w_{hi} (\bar{\mathbf{x}}_{whi} - \bar{\mathbf{x}}_h)^T \mathbf{d}_h + \delta_i. \end{aligned} \quad (11)$$

Typically, D will be large and the leading term in the variance of (9) will be the variance of δ_i in (11). If this leading term is large, then $\hat{V}\{\hat{B}(\hat{m}_i)\}$ will also be large, and (10) could take negative values. We therefore recommend that (8), rather than (10), be used. An immediate consequence is that (8) is then a conservative estimator of the MSE of \hat{m}_i under (1). This may be acceptable provided that the variance of δ_i is small. However, for very small values of n_i this variance can be large, causing (8) to substantially overestimate the actual MSE of \hat{m}_i . We therefore recommend a preliminary empirical assessment of the size of the variance of δ_i relative to the value of (7) in this situation. If this assessment indicates that the variance of δ_i dominates (7), then (8) should not be used.

2.2 MSE estimation for pseudo-linear small area estimators

The approach to conditional MSE estimation outlined in the previous sub-section assumed that the weights defining the linear estimator \hat{m}_i do not depend on the sample values of Y . However, most small area estimators do not satisfy this condition, in the sense that they are pseudo-linear in structure, with weights that do depend on these sample values. For example, the Best Linear Unbiased Predictor (BLUP) of m_i under the linear mixed model variant of (1) where the area-specific regression parameters β_i are independent and identically distributed realisations of a random variable with expected value β and covariance matrix Γ , can be written as a weighted sum of the sample

values of Y where the weights depend on Γ (see Royall 1976). Consequently, the empirical version of this predictor, the widely used EBLUP, is computed by substituting an efficient sample estimate of Γ (e.g., the REML estimate) into the BLUP weights. If the linear mixed model assumption is true, this sample estimator of Γ converges to the true value and consequently the EBLUP weights converge to the BLUP weights. That is, for large values of the overall sample size n , we can treat the EBLUP weights as fixed and use the MSE estimator (8) for the EBLUP. Of course, the EBLUP weights are not really fixed, and so (8) is therefore an approximation to the true MSE of the EBLUP that ignores the contribution to this MSE arising from the variability in estimation of Γ . However, this potential underestimation needs to be balanced against the bias robustness of (8) under misspecification of the second order moments of Y .

An important advantage of (8) is that it can be used with a range of small area estimators that can be expressed in pseudo-linear form. In particular, many small area estimators developed under models that are variants of (1) can be written in this form, *i.e.*, as weighted sums of the sample values of Y . To illustrate, we now focus on three such estimators: the EBLUP (Rao 2003, Chapter 6), the Model-Based Direct Estimator (MBDE) of Chandra and Chambers (2009) and the M-quantile predictor of Chambers and Tzavidis (2006). Each of these estimators can be written in pseudo-linear form, with weights that satisfy $w_{ij} = O(n_i^{-1})$ for $j \in s_i$ and $w_{ij} = o(n_i^{-1})$ for $j \notin s_i$, and so (8) can be used.

2.2.1 MSE estimation for the EBLUP

We first consider the well-known EBLUP for m_i based on a unit level linear mixed model extension of (1) of the form

$$\mathbf{y}_i = \mathbf{X}_i \beta + \mathbf{Z}_i \mathbf{u}_i + \mathbf{e}_i \quad (12)$$

where \mathbf{y}_i is the N_i -vector of population values of y_j in area i , \mathbf{X}_i is the corresponding $N_i \times p$ matrix of auxiliary variable values \mathbf{x}_j , \mathbf{Z}_i is the $N_i \times q$ component of \mathbf{X}_i corresponding to the q random components of β , \mathbf{u}_i is the associated q -vector of area-specific random effects and \mathbf{e}_i is the N_i -vector of individual random effects. It is typically assumed that the area and individual effects are mutually independent, with the area effects independently and identically distributed as $N(0, \Omega)$ and the individual effects independently and identically distributed as $N(0, \sigma^2)$. See Rao (2003, Chapter 6) for development of the underlying theory of this predictor. We note that the EBLUP can be written in pseudo-linear form,

$$\hat{m}_i^{\text{EBLUP}} = \sum_{j \in s} w_{ij}^{\text{EBLUP}} y_j = (\mathbf{w}_{is}^{\text{EBLUP}})^T \mathbf{y}_s \quad (13)$$

where

$$\begin{aligned} \mathbf{w}_{is}^{\text{EBLUP}} &= (\mathbf{w}_{ij}^{\text{EBLUP}}) \\ &= N_i^{-1} [\Delta_{is} + \{\hat{\mathbf{H}}_s^T \mathbf{X}_r^T + (\mathbf{I}_n - \hat{\mathbf{H}}_s^T \mathbf{X}_s^T) \hat{\Sigma}_{ss}^{-1} \hat{\Sigma}_{sr}\} \Delta_{ir}]. \end{aligned}$$

Here Δ_{ir} is the vector of size $N - n$ that ‘picks out’ the non-sampled units in area i , \mathbf{X}_s and \mathbf{X}_r are the matrices of order $n \times p$ and $(N - n) \times p$ respectively of the sample and non-sample values of the auxiliary variables, \mathbf{I}_n is the identity matrix of order n , $\hat{\mathbf{H}}_s = (\mathbf{X}_s^T \hat{\Sigma}_{ss}^{-1} \mathbf{X}_s)^{-1} \mathbf{X}_s^T \hat{\Sigma}_{ss}^{-1}$, $\hat{\Sigma}_{ss} = \hat{\sigma}^2 \mathbf{I}_n + \text{diag}\{\mathbf{Z}_{is}^T \hat{\Omega} \mathbf{Z}_{is}; i=1, \dots, D\}$ and $\hat{\Sigma}_{sr} = \text{diag}\{\mathbf{Z}_{is}^T \hat{\Omega} \mathbf{Z}_{ir}; i=1, \dots, D\}$. Here \mathbf{Z}_{is} (\mathbf{Z}_{ir}) is the sample (non-sample) component of \mathbf{Z}_i and $\hat{\sigma}^2$ and $\hat{\Omega}$ are suitable (e.g., ML or REML) estimates of the variance components of (12).

Given this setup, estimation of the conditional MSE of the EBLUP can be carried out using (8) with weights defined following (13). In turn, this requires that we have access to unbiased estimators $\hat{\mu}_j$ of the area specific individual expected values μ_j . However, such estimators may be unstable when area sample sizes are small. Consequently, it is tempting to replace $\hat{\mu}_j$ by the EBLUP for y_j , i.e., $\hat{y}_j^{\text{EBLUP}} = \mathbf{x}_j^T \hat{\beta}^{\text{EBLUE}} + \mathbf{z}_j^T \hat{\mathbf{u}}_i^{\text{EBLUP}}$, where $\hat{\beta}^{\text{EBLUE}}$ denotes the Empirical Best Linear Unbiased Estimator of β in the linear mixed model (12) and $\hat{\mathbf{u}}_i^{\text{EBLUP}}$ denotes the predicted area effect for the area i that contains observation j . Unfortunately, because of the well-known shrinkage effect associated with EBLUPs, this approach is not recommended. To illustrate this, we note that $\hat{V}(\hat{m}_i)$ in (8) uses $(y_j - \hat{\mu}_j)^2$ as an estimator of $E(y_j - \mu_j)^2$. The bias in this estimator is therefore

$$\begin{aligned} E(y_j - \hat{\mu}_j)^2 - E(y_j - \mu_j)^2 \\ &= -2E(y_j - \mu_j)(\hat{\mu}_j - \mu_j) + E(\hat{\mu}_j - \mu_j)^2 \\ &= -E\{(\hat{\mu}_j - \mu_j)(2y_j - \mu_j - \hat{\mu}_j)\} \end{aligned}$$

so we anticipate that $\hat{V}(\hat{m}_i)$ will be negatively biased if $E\{(\hat{\mu}_j - \mu_j)(2y_j - \mu_j - \hat{\mu}_j)\}$ is positive and vice versa. Now let sample unit j be from area i and consider the special case of a random intercept model for y_j , i.e., $y_j = \mathbf{x}_j^T \beta + u_i + e_j$ where u_i is the random effect for area i and e_j is a random individual effect uncorrelated with u_i . Here $\mu_j = \mathbf{x}_j^T \beta + u_i$. Suppose that we have a large overall sample size, allowing us to replace $\hat{\beta}^{\text{EBLUE}}$ by β . The EBLUP $\hat{\mu}_j = \hat{y}_j^{\text{EBLUP}}$ can then be approximated by $\tilde{\mu}_j = \mathbf{x}_j^T \beta + \gamma_i u_i$, where γ_i is a ‘shrinkage’ factor. It follows that

$$(\tilde{\mu}_j - \mu_j)(2y_j - \mu_j - \tilde{\mu}_j) = 2u_i(\gamma_i - 1)e_j - u_i^2(\gamma_i - 1)^2$$

so $E(y_j - \hat{\mu}_j)^2 - E(y_j - \mu_j)^2 \approx (\gamma_i - 1)^2 \sigma_u^2$. That is, we expect $\hat{V}(\hat{m}_i)$ to be positively biased if we use the shrunken

EBLUP \hat{y}_j^{EBLUP} to define $\hat{\mu}_j$. We also note that this bias disappears (approximately) if we ‘unshrink’ the residual component of this EBLUP. For example, in the case of the popular random intercepts model, we use

$$\hat{\mu}_j = \mathbf{x}_j^T \hat{\beta}^{\text{EBLUE}} + (\bar{y}_{is} - \bar{\mathbf{x}}_{is}^T \hat{\beta}^{\text{EBLUE}}) = \bar{y}_{is} + (\mathbf{x}_j - \bar{\mathbf{x}}_{is})^T \hat{\beta}^{\text{EBLUE}}$$

where \bar{y}_{is} and $\bar{\mathbf{x}}_{is}$ denote the sample means of Y and X respectively in area i . Given (12) is the working model, a general expression for such an ‘unshrunk’ estimator is

$$\hat{\mu}_j = \mathbf{x}_j^T \hat{\beta}^{\text{EBLUE}} + \mathbf{z}_j^T \tilde{\mathbf{u}}_i \quad (14)$$

where $\tilde{\mathbf{u}}_i = (\mathbf{Z}_{is}^T \mathbf{Z}_{is})^{-1} \mathbf{Z}_{is}^T (y_{is} - \mathbf{X}_{is} \hat{\beta}^{\text{EBLUE}})$ is the unshrunk predictor of the random effect for area i . It is not difficult to see that then $\hat{\mu}_j = \sum_{k \in s} \phi_{kj} y_k$ where $\phi_{kj} = c_{ijsk} + b_{ijsk} I(k \in i)$, with

$$\begin{aligned} c_{ijs} &= (c_{ijsk}; k \in s) \\ &= \hat{\Sigma}_{ss}^{-1} \mathbf{X}_s (\mathbf{X}_s^T \hat{\Sigma}_{ss}^{-1} \mathbf{X}_s)^{-1} \{\mathbf{x}_j - \mathbf{X}_{is}^T \mathbf{Z}_{is} (\mathbf{Z}_{is}^T \mathbf{Z}_{is})^{-1} \mathbf{z}_j\} \end{aligned}$$

and $b_{ijs} = (b_{ijsk}; k \in s_i) = \mathbf{Z}_{is} (\mathbf{Z}_{is}^T \mathbf{Z}_{is})^{-1} \mathbf{z}_j$. Note that these ϕ_{kj} ’s are also used to calculate the value of $\hat{\lambda}_j$ defined immediately after (7).

Finally, we observe that when (14) is used in (8), the estimated bias (9) becomes

$$\hat{B}(\hat{m}_i) = \sum_{h=1}^D \left(\sum_{j \in s_h} \mathbf{w}_{ij}^{\text{EBLUP}} \mathbf{z}_j \right)^T \tilde{\mathbf{u}}_h - \bar{\mathbf{z}}_i^T \tilde{\mathbf{u}}_i$$

since the EBLUP weights (13) are ‘locally calibrated’ on X , i.e., $\sum_{j \in s} \mathbf{w}_{ij}^{\text{EBLUP}} \mathbf{x}_j = \bar{\mathbf{x}}_i$. It follows that in this case the variable δ_i defined immediately before (11) takes the form

$$\delta_i = \sum_{h=1}^D \mathbf{w}_{hi}^{\text{EBLUP}} \bar{\mathbf{z}}_h^T \tilde{\mathbf{u}}_h - \bar{\mathbf{z}}_i^T \tilde{\mathbf{u}}_i$$

where $\mathbf{w}_{hi}^{\text{EBLUP}} = \sum_{j \in s_h} \mathbf{w}_{ij}^{\text{EBLUP}}$. For a large enough overall sample size δ_i can be approximated by

$$\begin{aligned} \delta_i &\approx \sum_{h=1}^D \mathbf{w}_{hi}^{\text{BLUP}} \bar{\mathbf{z}}_h^T (\mathbf{Z}_{hs}^T \mathbf{Z}_{hs})^{-1} \mathbf{Z}_{hs}^T (y_{hs} - \mathbf{X}_{hs} \beta) \\ &\quad - \bar{\mathbf{z}}_i^T (\mathbf{Z}_{is}^T \mathbf{Z}_{is})^{-1} \mathbf{Z}_{is}^T (y_{is} - \mathbf{X}_{is} \beta) \\ &= \sum_{h=1, h \neq i}^D \mathbf{w}_{hi}^{\text{BLUP}} \bar{\mathbf{z}}_h^T \{\mathbf{u}_h + (\mathbf{Z}_{hs}^T \mathbf{Z}_{hs})^{-1} \mathbf{Z}_{hs}^T \mathbf{e}_{hs}\} \end{aligned}$$

where $\mathbf{w}_{hi}^{\text{BLUP}}$ is the BLUP equivalent of $\mathbf{w}_{hi}^{\text{EBLUP}}$. The variance of δ_i can therefore be estimated via

$$\hat{V}(\delta_i) = \sum_{h=1, h \neq i}^D (\mathbf{w}_{hi}^{\text{EBLUP}})^2 \bar{\mathbf{z}}_h^T \{\hat{\Omega} + \hat{\sigma}^2 (\mathbf{Z}_{hs}^T \mathbf{Z}_{hs})^{-1}\} \bar{\mathbf{z}}_h. \quad (15)$$

If $\hat{V}(\delta_i)$ is small relative to the value of (7) in this case, then (8) can be used to estimate the MSE of the EBLUP. However, when n_i is very small, this condition may not hold. In such cases it may be advisable to consider more

model-dependent MSE estimators like the Prasad-Rao (PR) MSE estimator (Prasad and Rao 1990; Rao 2003, section 7.2.3). When a random means model is assumed, but the between area variability is very small relative to the within area variability, this advice extends to moderate area sample sizes as we now show.

2.2.2 MSE estimation for the EBLUP under the random means model

The random means model is the special case of (12) where $y_j = \beta + u_i + e_j$, with $u_i \sim N(0, \sigma_u^2)$ and $e_j \sim N(0, \sigma^2)$. The EBLUE of β is then $\hat{\beta} = \sum_{h=1}^D \hat{\alpha}_h \bar{y}_{hs}$ with $\hat{\alpha}_i = (\hat{\phi} + n_i^{-1})^{-1} \{ \sum_{h=1}^D (\hat{\phi} + n_h^{-1})^{-1} \}^{-1}$ and $\hat{\phi} = \hat{\sigma}_u^2 / \hat{\sigma}^2$, and the EBLUP (13) is defined by weights of the form

$$w_{ij}^{\text{EBLUP}} = (1 - f_i)(1 - \hat{\gamma}_i) \sum_{h=1}^D \hat{\alpha}_h n_h^{-1} I(j \in h) \\ + \{f_i + (1 - f_i) \hat{\gamma}_i\} n_i^{-1} I(j \in i)$$

with $\hat{\gamma}_i = n_i \hat{\phi} (1 + n_i \hat{\phi})^{-1}$. For $j \in h$, $\hat{\mu}_j = \sum_{k \in s} \phi_{kj} y_k = \bar{y}_{hs}$ and so

$$\hat{\lambda}_j = (1 - \phi_{jj})^2 + \sum_{k \in s(-j)} \phi_{kj}^2 \\ = (1 - n_h^{-1})^2 + (n_h - 1) n_h^{-2} = (n_h - 1) n_h^{-1}.$$

It follows that the estimator (7) of the conditional prediction variance of \hat{m}_i^{EBLUP} in this case is

$$\hat{V}(\hat{m}_i^{\text{EBLUP}}) = (1 - f_i)^2 \left[\sum_{h=1}^D \{ (1 - \hat{\gamma}_i)^2 \hat{\alpha}_h^2 n_h^{-2} \right. \\ \left. + (N_i - n_i)^{-1} n^{-1} \} n_h s_h^2 \right. \\ \left. + \hat{\gamma}_i n_i^{-1} \{ 2(1 - \hat{\gamma}_i) \hat{\alpha}_i + \hat{\gamma}_i \} s_i^2 \right],$$

where $s_h^2 = (n_h - 1)^{-1} \sum_{j \in s_h} (y_j - \bar{y}_{hs})^2$, while from (9) the estimator of the conditional prediction bias of \hat{m}_i^{EBLUP} is $\hat{B}(\hat{m}_i^{\text{EBLUP}}) = (1 - f_i)(1 - \hat{\gamma}_i)(\hat{\beta} - \bar{y}_{is})$. For $h \neq i$ we also then have

$$w_{hi}^{\text{EBLUP}} = \sum_{j \in s_h} w_{ij}^{\text{EBLUP}} \\ = (1 - f_i) \hat{\alpha}_h (1 + n_i \hat{\phi})^{-1} \approx \hat{\alpha}_h (1 + n_i \hat{\phi})^{-1}$$

when we ignore $O(N_i^{-1})$ terms. A similar approximation to (15) therefore leads to

$$\hat{V}(\delta_i) = \sum_{h=1}^D (w_{hi}^{\text{EBLUP}})^2 (\hat{\sigma}_u^2 + n_h^{-1} \hat{\sigma}^2) \\ \approx \hat{\sigma}^2 \sum_{h=1}^D \left(\frac{\hat{\alpha}_h}{1 + n_i \hat{\phi}} \right)^2 \left(\frac{1 + n_h \hat{\phi}}{n_h} \right).$$

Suppose now that the sample size in every small area is the same, i.e., $n_i = m$. Then $n = mD$, $\hat{\alpha}_h = D^{-1}$ and the approximation to $\hat{V}(\delta_i)$ above takes the form

$$\hat{V}(\delta_i) = \hat{\sigma}^2 \sum_{h=1}^D \left(\frac{D^{-1}}{1 + m \hat{\phi}} \right)^2 \left(\frac{1 + m \hat{\phi}}{m} \right) \approx n^{-1} (1 + m \hat{\phi})^{-1} \hat{\sigma}^2$$

while the corresponding approximation to $\hat{V}(\hat{m}_i^{\text{EBLUP}})$ is

$$\hat{V}(\hat{m}_i^{\text{EBLUP}}) \approx \sum_{h=1}^D (1 + m \hat{\phi})^{-2} D^{-2} m^{-1} s_h^2 \\ + (1 + m \hat{\phi})^{-2} \hat{\phi} (2D^{-1} + m \hat{\phi}) s_i^2 \\ = n^{-1} (1 + m \hat{\phi})^{-2} \left\{ \left(D^{-1} \sum_{h=1}^D s_h^2 \right) + m \hat{\phi} (2 + n \hat{\phi}) s_i^2 \right\}.$$

Comparing these approximations to $\hat{V}(\delta_i)$ and $\hat{V}(\hat{m}_i^{\text{EBLUP}})$ we see that if $m \hat{\phi}$ is small (e.g., when m and $\hat{\phi}$ are both small) then these terms will be of similar magnitude. In this situation we expect (8) to overestimate the true MSE of the EBLUP. In particular, the approximation to (8) when $m \hat{\phi}$ is small and N_i is large is

$$\hat{M}(\hat{m}_i^{\text{EBLUP}}) \approx n^{-1} \left(D^{-1} \sum_{h=1}^D s_h^2 \right) + \left(\bar{y}_{is} - D^{-1} \sum_{h=1}^D \bar{y}_{hs} \right)^2. \quad (16)$$

Note that the expectation of the squared residual on the right hand side of (16) when $m \hat{\phi}$ is small is $(1 - D^{-1})(\sigma_u^2 + m^{-1} \sigma^2) = O(1)$ and so it is the leading term in this estimator in this situation. This expression can be compared with the corresponding one for the MSE estimator of the EBLUP suggested by Prasad and Rao (1990). Under the random means model, the PR MSE estimator is

$$\hat{M}_{\text{PR}}(\hat{m}_i^{\text{EBLUP}}) = (1 - f_i)^2 \left[\hat{\gamma}_i m^{-1} \hat{\sigma}^2 \right. \\ \left. + (1 - \hat{\gamma}_i)^2 \left(m \sum_{h=1}^D \hat{\tau}_h^{-1} \right)^{-1} + N_i^{-1} (1 - f_i)^{-1} \hat{\sigma}^2 \right. \\ \left. + \frac{2}{T} m \hat{\tau}_i^{-3} \left\{ \hat{\sigma}^4 \left(\frac{n - D}{\hat{\sigma}^4} + \sum_{h=1}^D \hat{\tau}_h^{-2} \right) \right. \right. \\ \left. \left. + \hat{\sigma}_u^4 m^2 \sum_{h=1}^D \hat{\tau}_h^{-2} + 2 \hat{\sigma}^2 \hat{\sigma}_u^2 m \sum_{h=1}^D \hat{\tau}_h^{-2} \right\} \right]$$

where $\hat{\tau}_i = n_i \hat{\sigma}_u^2 + \hat{\sigma}^2$ and

$$T = \frac{n - D}{\hat{\sigma}^4} \sum_{h=1}^D n_h^2 \hat{\tau}_h^{-2} \\ + \left(\sum_{h=1}^D \hat{\tau}_h^{-2} \right) \left(\sum_{h=1}^D n_h^2 \hat{\tau}_h^{-2} \right) - \left(\sum_{h=1}^D n_h \hat{\tau}_h^{-2} \right)^2.$$

Assuming $n_i = m$, $m \hat{\phi}$ is small and N_i is large, $\hat{M}_{\text{PR}}(\hat{m}_i^{\text{EBLUP}})$ has the approximation

$$\hat{M}_{PR}(\hat{m}_i^{EBLUP}) \approx \hat{\sigma}^2 \{n^{-1} + 2(n-D)^{-1}\} + \hat{\sigma}_u^2. \quad (17)$$

Comparing (16) and (17) we can see that the instability and the overestimation associated with the use of (8) in this situation are both due to the use of the square of the single degree of freedom area level residual $\bar{y}_{is} - D^{-1} \sum_{h=1}^D \bar{y}_{hs}$ as an estimator of σ_u^2 . This reinforces earlier comments that (8) should not generally be used for estimating the MSE of the EBLUP if the area sample sizes are very small or, in the special case of the random means model, for moderate area sample sizes when the between area variability is very small relative to the within area variability.

2.2.3 MSE estimation for the MBDE

The second predictor of m_i that we consider is the Model-Based Direct Estimator (MBDE) described in Chandra and Chambers (2009). This is based on the same linear mixed model (12) as the EBLUP, with the MBDE predictor defined as

$$\hat{m}_i^{MBDE} = \sum_{j \in s} w_{ij}^{MBDE} y_j = (\mathbf{w}_{is}^{MBDE})^T \mathbf{y}_s \quad (18)$$

where

$$w_{ij}^{MBDE} = \frac{I(j \in s_i) w_j^{EBLUP}}{\sum_{k \in s} I(k \in s_i) w_k^{EBLUP}}. \quad (19)$$

Here $I(j \in s_i)$ is the indicator function for unit j to be in the area i sample, and $\mathbf{w}_s^{EBLUP} = (w_j^{EBLUP})$ is the vector of weights that defines the EBLUP for the population total of the y_j under (12), i.e.,

$$\mathbf{w}_s^{EBLUP} = (\mathbf{w}_j^{EBLUP}) = \mathbf{1}_n + \{\hat{\mathbf{H}}_s^T \mathbf{X}_r^T + (\mathbf{I}_n - \hat{\mathbf{H}}_s^T \mathbf{X}_s^T) \hat{\Sigma}_{ss}^{-1} \hat{\Sigma}_{sr}\} \mathbf{1}_{N-n}$$

where $\mathbf{1}_n$ ($\mathbf{1}_{N-n}$) denotes the unit vector of size n ($N-n$) and $\hat{\mathbf{H}}_s$ was defined in section 2.2.1. In this case pseudo-linearisation based estimation of the area-specific MSE of the MBDE is carried out using (8), with weights defined by (19). Note that the estimated expected values used in (8) when applied to the MBDE are the same as the unshrunk estimates (14) used with the EBLUP, reflecting the fact that both the MBDE and the EBLUP are based on the same linear mixed model (12). However, the MBDE weights (19) are not locally calibrated, and so the squared bias term in (8) cannot be ignored when estimating the MSE of this predictor. Furthermore, since

$$w_{hi}^{MBDE} = \sum_{j \in s_h} w_{ij}^{MBDE} = 0$$

for $h \neq i$, we have $\delta_i = 0$ for the MBDE and so the bias estimator (9) works well in this case.

2.2.4 MSE estimation for the M-quantile estimator

The third estimator that we consider is based on the M-quantile modelling approach described in Chambers and Tzavidis (2006). This approach does not assume an underlying linear mixed model, relying instead on characterising the relationship between y_j and \mathbf{x}_j in area i in terms of the linear M-quantile model that best 'fits' the sample y_j values from this area. That is, this approach replaces (12) by a model of the form

$$y_i = \mathbf{X}_i \boldsymbol{\beta}(q_i) + \mathbf{e}_i \quad (20)$$

where $\boldsymbol{\beta}(q)$ denotes the coefficient vector of a linear model for the regression M-quantile of order q for the population values of Y and X , and q_i denotes the M-quantile coefficient of area i . Given an estimate \hat{q}_i of q_i , an iteratively re-weighted least squares (IRLS) algorithm is used to calculate an estimate

$$\hat{\boldsymbol{\beta}}(\hat{q}_i) = \{\mathbf{X}_s' \mathbf{W}_s(\hat{q}_i) \mathbf{X}_s\}^{-1} \mathbf{X}_s' \mathbf{W}_s(\hat{q}_i) \mathbf{y}_s \quad (21)$$

of $\boldsymbol{\beta}(q_i)$ in (20), and a non-sample value of y_j in area i is then predicted by $\hat{y}_j = \mathbf{x}_j^T \hat{\boldsymbol{\beta}}(\hat{q}_i)$. Here $\mathbf{W}_s(\hat{q}_i)$ is the diagonal matrix of final weights used in the IRLS algorithm.

Tzavidis, Marchetti and Chambers (2010) note that value of the M-quantile estimator suggested in Chambers and Tzavidis (2006) can be interpreted as the expected value of Y in area i with respect to a biased estimator of the distribution function of this variable in the area. They therefore develop an improved M-quantile estimator, replacing this biased distribution function estimator by the Chambers and Dunstan (1986) distribution function estimator under the area-specific model (1). This corresponds to predicting m_i by

$$\hat{m}_i^{MQ} = \sum_{j \in s} w_{ij}^{MQ} y_j = (\mathbf{w}_{is}^{MQ})^T \mathbf{y}_s \quad (22)$$

where

$$\mathbf{w}_{is}^{MQ} = n_i^{-1} \Delta_{is} + (1 - N_i^{-1} n_i) \mathbf{W}_s(\hat{q}_i) \mathbf{X}_s \{\mathbf{X}_s^T \mathbf{W}_s(\hat{q}_i) \mathbf{X}_s\}^{-1} (\bar{\mathbf{x}}_{is} - \bar{\mathbf{x}}_{is}).$$

Here $\bar{\mathbf{x}}_{is}$ and $\bar{\mathbf{x}}_{ir}$ are the vectors of sample and non-sample means of the x_j in area i . It is not difficult to show that the weights following (22) are locally calibrated. Furthermore, if we then put $\hat{\mu}_j = \mathbf{x}_j^T \hat{\boldsymbol{\beta}}(\hat{q}_i)$, where $\hat{\boldsymbol{\beta}}(\hat{q}_i)$ is defined by (21), it is easy to see that (9) is zero and so the area-specific MSE of the bias-corrected M-quantile estimator (22) can be estimated using just the estimated prediction variance component (7). Since the constant $\hat{\lambda}_j$ in (7) is typically very close to one under M-quantile estimation, we set it equal to this value whenever we compute values of (7) that relate to

small area estimation (SAE) under the M-quantile modelling approach.

As we have already done with the EBLUP, we note that use of (7) implicitly treats the weights defining (22) as fixed, which is actually not the case since the matrix $\mathbf{W}_s(\hat{q}_i)$ is a function of the sample values of Y . An immediate consequence is that pseudo-linearisation based estimation of the MSE of the M-quantile predictor via (7) is a first order approximation to the true MSE of this estimator. Nevertheless, since accounting for weight variability in the definition of the M-quantile estimator considerably complicates estimation of its MSE - see Street, Carroll and Ruppert (1988) for an examination of this issue in the context of 'standard' M-estimation of regression coefficients - it is of interest to see how the relatively simple estimator (7) performs when used to estimate this MSE.

2.3 MSE estimation for the pseudo-linear synthetic EBLUP

In many SAE applications there are areas that contain no sample, and hence synthetic estimation is used. Although such estimators do not fit into the class of pseudo-linear estimators considered in this paper, the ideas behind the conditional MSE estimator (8) can be applied here as well. To see this, assume that these areas are numbered last, *i.e.*, if D^+ areas have non-zero sample then $n_h > 0$ for $h \leq D^+$ and $n_h = 0$ for $h > D^+$. For $i > D^+$ the 'synthetic EBLUP' for m_i is

$$\begin{aligned}\hat{m}_i^{\text{SYN-EBLUP}} &= \bar{\mathbf{x}}_i^T \hat{\boldsymbol{\beta}}^{\text{EBLUE}} = (\mathbf{w}_{is}^{\text{SYN-EBLUP}})^T \mathbf{y}_s \\ &= \sum_{h=1}^{D^+} \sum_{j \in s_h} w_{ij}^{\text{SYN-EBLUP}} y_j\end{aligned}\quad (23)$$

where

$$\mathbf{w}_{is}^{\text{SYN-EBLUP}} = (\mathbf{w}_{ij}^{\text{SYN-EBLUP}}) = \hat{\mathbf{H}}_s^T \bar{\mathbf{x}}_i.$$

Clearly (23) is a pseudo-linear estimator, and so we can use (7) to estimate its prediction variance, observing that since $n_i = 0$, $a_{ij} = N_i w_{ij}^{\text{EBLUP}}$ and so (7) becomes

$$\begin{aligned}\hat{V}(\hat{m}_i^{\text{SYN-EBLUP}}) &= \\ &\sum_{j \in s} \{ (w_{ij}^{\text{SYN-EBLUP}})^2 + N_i^{-1} n^{-1} \} \hat{\lambda}_j^{-1} (y_j - \hat{\mu}_j)^2.\end{aligned}\quad (24)$$

Unfortunately, since there is no sample in area i , we cannot use (9) to estimate the area-specific bias (2) of $\hat{m}_i^{\text{SYN-EBLUP}}$. However, under the linear mixed model (12), this bias has expected value

$$\begin{aligned}E(\hat{m}_i^{\text{SYN-EBLUP}} - m_i) &= \\ &\sum_{h=1}^{D^+} \sum_{j \in s_h} w_{ij}^{\text{SYN-EBLUP}} (\mathbf{x}_j^T \boldsymbol{\beta} + \mathbf{z}_j^T \mathbf{u}_h) - \bar{\mathbf{x}}_i^T \boldsymbol{\beta} - \bar{\mathbf{z}}_i^T \mathbf{u}_i.\end{aligned}$$

The conditional expectation of the square of this expected bias, given the area effects $\mathbf{u}_s = (\mathbf{u}_h; h = 1, \dots, D^+)$ for the sampled areas, is

$$\begin{aligned}E\{E^2(\hat{m}_i^{\text{SYN-EBLUP}} - m_i) | \mathbf{X}, \mathbf{u}_s\} &= \\ &\left\{ \sum_{h=1}^{D^+} \sum_{j \in s_h} w_{ij}^{\text{SYN-EBLUP}} (\mathbf{x}_j^T \boldsymbol{\beta} + \mathbf{z}_j^T \mathbf{u}_h) - \bar{\mathbf{x}}_i^T \boldsymbol{\beta} \right\}^2 + \bar{\mathbf{z}}_i^T \boldsymbol{\Omega} \bar{\mathbf{z}}_i,\end{aligned}$$

which immediately suggests that for a non-sampled area i we estimate the squared bias of the synthetic estimator $\hat{m}_i^{\text{SYN-EBLUP}}$ by

$$\begin{aligned}\hat{B}^2(\hat{m}_i^{\text{SYN-EBLUP}}) &= \\ &\left\{ \sum_{h=1}^{D^+} \sum_{j \in s_h} w_{ij}^{\text{SYN-EBLUP}} (\mathbf{x}_j^T \hat{\boldsymbol{\beta}}^{\text{EBLUE}} + \mathbf{z}_j^T \tilde{\mathbf{u}}_h) \right. \\ &\quad \left. - \bar{\mathbf{x}}_i^T \hat{\boldsymbol{\beta}}^{\text{EBLUE}} \right\}^2 + \bar{\mathbf{z}}_i^T \hat{\boldsymbol{\Omega}} \bar{\mathbf{z}}_i.\end{aligned}\quad (25)$$

Here $\tilde{\mathbf{u}}_h$ is the 'unshrunk' estimated effect for sampled area h - see (14). Our proposed MSE estimator for $\hat{m}_i^{\text{SYN-EBLUP}}$ is then the sum of (24) and (25). Note that, unlike (8), this MSE estimator includes no information from area i , and so is not an estimator of the area-specific MSE of (23). In particular, its validity depends completely on the mixed model (12) holding, and so it is not robust to misspecification of this model.

3. Simulation studies of the proposed MSE estimator

In this section we describe results from five simulation studies that aim at assessing the performance of the approach to conditional MSE estimation described in the previous section. Three of these studies are model-based simulations, with population data generated from the linear mixed model (12). The remaining two are design-based simulations, with population data derived from two real survey datasets where linear SAE is of interest.

Given our focus on bias-robustness, the main performance indicator for an MSE estimator in all five studies is its median relative bias, defined by

$$\text{RB}(M) = \text{median}_i \left\{ M_i^{-1} K^{-1} \sum_{k=1}^K (\hat{M}_{ik} - M_i) \right\} \times 100.$$

Here the subscript i indexes the small areas and the subscript k indexes the K Monte Carlo simulations, with \hat{M}_{ik} denoting the simulation k value of the MSE estimator in area i , and M_i denotes the actual (*i.e.*, Monte Carlo) MSE in area i . Since we would naturally prefer to use the more stable of two approximately unbiased MSE estimators, we also

measured the stability of an MSE estimator by its median percent relative root mean squared error,

$$\text{RRMSE}(M) = \text{median}_i \left\{ \sqrt{K^{-1} \sum_{k=1}^K \left(\frac{\hat{M}_{ik} - M_i}{M_i} \right)^2} \right\} \times 100.$$

Although the purpose of this paper is not to compare different methods of SAE, it is useful to relate MSE estimation performance for a particular method of SAE to the actual estimation performance of this method. We therefore provide two measures of the relative performance of the SAE methods that were used in our simulations. These are the median percent relative bias

$$\text{RB}(m) = \text{median}_i \left\{ \bar{m}_i^{-1} K^{-1} \sum_{k=1}^K (\hat{m}_{ik} - m_{ik}) \right\} \times 100$$

and the median percent relative root mean squared error

$$\text{RRMSE}(m) = \text{median}_i \left\{ \sqrt{K^{-1} \sum_{k=1}^K \left(\frac{\hat{m}_{ik} - m_{ik}}{m_{ik}} \right)^2} \right\} \times 100$$

of the estimates \hat{m}_{ik} generated by an estimation method. Note that $\bar{m}_i = K^{-1} \sum_{k=1}^K m_{ik}$ here.

3.1 Model-based simulations

The first model-based simulation study was based on population data generated under the mixed model (12) with Gaussian random effects. It used a population size of $N = 15,000$, with $D = 30$ small areas. Population sizes in the small areas were uniformly distributed over the interval [443, 542] and kept fixed over simulations. At each simulation, population values for Y were generated under the random intercepts model $y_j = 500 + 1.5x_j + u_i + e_j$, with x_j drawn from a chi-squared distribution with 20 degrees of freedom. The area effects u_i and individual

effects e_j were independently drawn from $N(0, \sigma_u^2)$ and $N(0, \sigma^2)$ distributions respectively, with the values of σ_u and σ shown in rows SIM1-A and SIM1-B of Table 1. A sample of size $n = 600$ was selected from each simulated population, with area sample sizes proportional to the fixed area populations, resulting in a median area sample size of $n_i = 20$. Sampling was via stratified random sampling, with the strata defined by the small areas. A total of $K = 1,000$ simulations were carried out.

Conditions for the second model-based simulation study were the same as in the first, with the exception that the area level random effects and the individual level random effects were independently drawn from mean corrected chi-square distributions respectively. The corresponding values of the area level and individual level variances are shown in rows SIM2-A and SIM2-B in Table 1. Finally, in the third model-based simulation study conditions were kept the same as in SIM1-A and SIM1-B for areas 1-25, but in areas 26-30 the area effects were independently drawn from a normal distribution with a larger variance. We refer to this as a Mixture in Table 1, with variances for areas 1-25 shown in rows SIM3-A and SIM3-B, and variances for areas 26-30 shown in rows SIM3-A* and SIM3B*. Our objective in this third simulation was to investigate the behaviour of the different methods of MSE estimation for 'outlier' areas, and so we show values relating to areas 1-25 and 26-30 separately in Tables 2 and 4. We also replicated all three scenarios above using a reduced overall sample size of $n = 150$ (with median area sample size $n_i = 5$). These additional simulations allowed us to investigate the effect of reduced sample sizes on the performance of the MSE estimators.

Table 1
Parameter values used in model-based simulations

Type	Simulation	σ_u^2	σ^2	$\rho = \sigma_u^2(\sigma_u^2 + \sigma^2)^{-1}$
Gaussian	SIM1-A	10.40	94.09	0.1
	SIM1-B	40.32	94.09	0.3
Chi-square	SIM2-A	2.0	10.0	0.1667
	SIM2-B	4.0	10.0	0.2857
Mixture (areas 1-25)	SIM3-A	10.40	94.09	0.10
	SIM3-B	40.32	94.09	0.30
Mixture (areas 26-30)	SIM3-A*	225.0	94.09	0.7051
	SIM3-B*	225.0	94.09	0.7051

Table 2
Median relative biases $RB(m)$ and median relative root mean squared errors $RRMSE(m)$ of estimators of small area means in model-based simulations

Weighting Method	Simulation							
	SIM1-A	SIM1-B	SIM2-A	SIM2-B	SIM3-A	SIM3-B	SIM3-A*	SIM3-B*
RB(m), median $n_i = 20$								
Regression	0.005	0.005	0.000	0.000	0.004	0.004	0.006	0.006
EBLUP, (13)	0.005	0.006	0.004	-0.002	0.004	0.005	0.006	0.005
MBDE, (18)	0.006	0.006	0.005	-0.008	0.007	0.007	0.001	0.001
M-quantile, (22)	0.009	0.008	-0.002	0.002	0.015	0.015	-0.013	-0.013
RRMSE(m), median $n_i = 20$								
Regression	0.40	0.40	0.13	0.13	0.40	0.40	0.41	0.41
EBLUP, (13)	0.35	0.38	0.12	0.13	0.37	0.38	0.45	0.42
MBDE, (18)	0.55	0.55	0.41	0.43	0.56	0.56	0.55	0.55
M-quantile, (22)	0.41	0.41	0.13	0.13	0.41	0.41	0.36	0.36
RB(m), median $n_i = 5$								
Regression	-0.002	-0.003	-0.001	0.002	-0.003	-0.004	0.011	0.011
EBLUP, (13)	0.001	0.005	-0.002	0.003	0.002	-0.001	0.008	0.011
MBDE, (18)	-0.002	-0.002	-0.005	0.004	-0.001	-0.002	-0.002	-0.002
M-quantile, (22)	-0.001	-0.001	-0.001	0.001	-0.003	-0.003	0.014	0.014
RRMSE(m), median $n_i = 5$								
Regression	0.81	0.81	0.26	0.26	0.82	0.82	0.80	0.80
EBLUP, (13)	0.53	0.69	0.19	0.22	0.61	0.71	1.00	0.87
MBDE, (18)	1.13	1.13	0.83	0.83	1.13	1.13	1.13	1.13
M-quantile, (22)	0.81	0.81	0.26	0.26	0.81	0.81	0.80	0.80

Table 2 shows the median bias $RB(m)$ and median relative root mean squared error $RRMSE(m)$ of the SAE methods investigated in our simulations for the two sample sizes ($n = 600$ and 150). These are the synthetic regression estimator (see Rao 2003, page 136), the EBLUP with weights defined by (13), the MBDE with weights defined by (18) and the M-quantile estimator defined by the weights (22). The differences between the various SAE estimators in Table 2 are essentially as one would expect. Bias is not really an issue (to be expected given the population data follow a linear model in all cases), while for Simulation scenarios 1 and 2 the indirect estimator (EBLUP) is the most efficient in terms of $RRMSE$. The M-quantile estimator is the best performer for SIM3-A* and SIM3-B* with $n_i = 20$ but its difference from the regression synthetic estimator reduces for the scenario with the smaller area-specific sample sizes. Note that in this case the M-quantile weights (22) are based on an outlier-robust estimate of the M-quantile coefficient \hat{q}_i for area i , defined by the median (rather than the mean) of the M-quantile coefficients of sampled units in this area. Further, as the sample sizes decrease, the $RRMSE$ s of all estimators increase, but their relative performances remain the same. Under normality the EBLUP is better than the M-quantile estimator but the differences between these two estimators become smaller as we move away from normality, with the M-quantile estimator more efficient in the mixture model scenarios.

Table 3 sets out the various MSE estimators investigated in our simulations that are based on the approach proposed in this paper. These are collectively referred to as “conditional”

MSE estimators below. In Table 4 we show the performances of MSE estimators for the small area estimators considered in Table 2. Note that in addition to the conditional MSE estimators, we provide results for three other MSE estimators for the EBLUP, with PR0 denoting the estimator suggested by Prasad and Rao (1990), see Rao (2003, section 6.2.6). It is noteworthy that PR0 is not an estimator of the area-specific MSE of the EBLUP, but of its MSE under the mixed linear model (12), *i.e.*, averaged over possible realisations of the area effect. In contrast, the MSE estimators PR1 and PR2 in Table 4 are the area-specific versions of PR0 suggested in Rao (2003, section 6.3.2 expressions 6.3.15 and 6.3.16 respectively). Finally, we note that the MSE estimator of the synthetic regression estimator that we used in our simulations is its variance estimator based on a fixed effects population regression model. We denote it by VReg.

The results set out in Table 4 focus on the median biases $RB(M)$ and median relative root mean squared error $RRMSE(M)$ of the various MSE estimators. Not surprisingly, given that all its underlying assumptions are met, the PR0 estimator and its area-specific alternatives, PR1 and PR2, perform very well in both normal scenarios (SIM1-A and SIM1-B) and both chi-squared scenarios (SIM2-A and SIM2-B), with virtually no bias ($n_i = 20$) or small bias when within area sample sizes are very small. For the MSE estimator of the synthetic regression estimator, on the other hand, we see substantial relative bias under all simulation scenarios.

Table 3
Definitions of conditional MSE estimators for different weighting methods

Weighting Method	Definition of $\hat{\mu}_j, j \in i$	MSE Estimator
EBLUP (13)	(14)	(8)
MBDE (18)	(14)	(8)
M-quantile (22)	$x_j^T \hat{\beta}(\hat{q}_i)$	(7) with $\hat{\lambda}_j = 1$
Synthetic EBLUP (23)	(14)	(24) + (25)

Table 4
Median relative biases RB(M) and median relative root mean squared errors RRMSE(M) for MSE estimators in model-based simulations

Weighting Method	MSE Estimator	Simulation							
		SIM1-A	SIM1-B	SIM2-A	SIM2-B	SIM3-A	SIM3-B	SIM3-A*	SIM3-B*
RB(M), median $n_i = 20$									
Regression	VReg	7.59	21.82	11.81	20.78	23.66	34.27	23.97	34.64
EBLUP, (13)	PR0	-0.83	-0.72	0.56	1.16	3.44	0.71	-15.65	-6.51
	PR1	-0.97	-0.72	0.64	1.08	2.94	0.56	-13.70	-5.81
	PR2	-0.92	-0.72	0.64	1.16	3.20	0.61	-14.65	-6.19
	Conditional	3.89	-0.89	3.06	0.93	-0.05	-0.54	-2.56	-1.59
MBDE, (18)	Conditional	-0.81	-0.80	-0.06	-0.42	-0.75	-0.75	-0.98	-0.98
M-quantile, (22)	Conditional	-3.10	-1.66	-0.09	-1.90	-5.04	-3.17	11.26	11.04
RRMSE(M), median $n_i = 20$									
Regression	VReg	18	51	30	53	59	85	60	86
EBLUP, (13)	PR0	12	7	15	10	11	7	29	14
	PR1	14	7	17	11	10	7	27	13
	PR2	12	7	16	10	11	7	28	13
	Conditional	62	31	70	49	31	30	42	32
MBDE, (18)	Conditional	70	70	126	128	71	71	67	67
M-quantile, (22)	Conditional	32	34	49	48	31	32	48	48
RB(M), median $n_i = 5$									
Regression	VReg	5.59	19.17	10.35	19.12	20.92	30.91	22.93	33.00
EBLUP, (13)	PR0	3.51	-0.20	2.42	1.19	12.79	3.86	-30.64	-15.92
	PR1	3.04	-0.50	2.13	1.00	10.84	3.10	-25.77	-13.62
	PR2	3.16	-0.31	2.31	1.11	11.81	3.48	-28.16	-14.77
	Conditional	37.52	4.38	24.11	8.93	8.18	1.50	-0.66	-0.68
MBDE, (18)	Conditional	-0.24	-0.21	0.02	-0.09	-0.62	-0.33	1.29	1.24
M-quantile, (22)	Conditional	-7.60	-6.17	5.70	5.00	-5.95	-5.60	5.89	3.60
RRMSE(M), median $n_i = 5$									
Regression	VReg	17	46	33	51	54	78	59	83
EBLUP, (13)	PR0	31	14	33	22	36	16	53	31
	PR1	48	18	44	28	34	16	48	29
	PR2	36	15	36	24	34	15	50	29
	Conditional	234	81	193	121	86	66	86	70
MBDE, (18)	Conditional	79	79	133	129	79	79	83	83
M-quantile, (22)	Conditional	62	63	90	97	63	63	122	102

The conditional MSE estimator for the EBLUP shows positive bias under both the normal (SIM1A) and chi-squared (SIM2A) scenarios, particularly for moderate intra-cluster correlation (3.89% and 37.52% for the normal scenario with 20 and 5 units in each area respectively and 3.06% and 24.11% for the chi-squared scenario with 20 and 5 units in each area respectively). This bias increases with decreasing sample size. However, things change when we

examine the results for the outlier components of the mixture model scenarios (SIM3-A* and SIM3-B*). Here we see a substantial negative bias for all three versions of PR (ranging from -30.64% to -5.81% depending on the area sample sizes). In comparison, the conditional MSE estimator for the EBLUP now shows a smaller negative bias (-2.56% and -0.66%) while the same MSE estimator applied to the M-quantile estimator shows an upward bias. The

conditional MSE estimator for the MBDE is essentially unbiased. Given that as far as MSE estimation is concerned, positive bias is preferable to negative bias, it seems clear that the proposed conditional MSE estimator is better able to handle this outlier situation. Figure 1 graphically illustrates this point for sample size $n = 600$. Here we show the area-specific RMSEs and the average (over the simulations) of the estimated RMSEs in each of the 30 areas for the mixture simulations SIM3-A and SIM3-A*, with the vertical line delineating the five 'outlier' areas. In the top panel of this plot we see that the PR0 estimator is unable to detect the step increase in the MSE of the EBLUP for these 'outlier' areas, being biased slightly high in the 'well-behaved' areas and then biased rather low in the 'outlier' areas. In contrast, the conditional MSE estimator for the EBLUP and the MBDE tracks the area specific RMSEs rather well, while the same MSE estimator based on M-quantile weights tends to be biased low in the 'well-behaved' areas, and biased high in the 'outlier' areas, which can be argued as being perhaps a rather better outcome than that recorded by the PR0 estimator in this simulation. It should be noted here that in certain circumstances an assumed model can be revised after outlier detection. However, this requires a sufficiently large number of detected outliers to permit their separate modelling. This is unlikely to happen in practice. Also, particular care must be taken with extrapolation of these results to the case of very small area sample sizes because of the instability that the conditional MSE estimator can exhibit in this case.

Table 4 also shows the relative RMSEs of the different MSE estimators across the three types of model-based simulation. Here we see that all three versions of the PR estimator of the MSE of the EBLUP are more stable than the conditional MSE estimator of the EBLUP (12% for PR vs. 62% for the conditional MSE for SIM1-A with $n_i = 20$ and 31% for PR vs. 234% for the conditional MSE for SIM1-A with $n_i = 5$). These differences decrease under scenarios SIM3-A* and SIM3-B*, however, although the PR MSE estimator remains more stable (13% for PR vs. 32% for the conditional MSE estimator for SIM3-B* with $n_i = 20$ and 29% for the PR MSE estimator vs. 70% for the conditional MSE estimator for SIM3-B* with $n_i = 5$). The same is true for the conditional MSE estimators of the MBDE and the M-quantile estimators. Essentially, given sample data that follow a mixed linear model, the PR MSE estimator of MSE is very stable, while the conditional MSE estimator is more variable.

In summary, although all methods of MSE estimation that we evaluated exhibited some bias for very small area sample sizes, our model-based simulation results provide evidence that for larger area sample sizes the conditional

MSE estimation method (8) is bias robust when applied to the three pseudo-linear small area estimators EBLUP, MBDE and M-quantile. For very small area sample sizes its bias robustness is less evident. As one might expect, the model dependent 'area-averaged' MSE estimator PR0 for the EBLUP exhibits bias under model failure. The fact that we observed rather similar behaviour for the area-specific versions PR1 and PR2 of this MSE estimator indicates that 'area specific' does not necessarily mean 'bias robust'. In particular, the fact that PR1 and PR2 behave very similarly to PR0 may be because the area-specific components of PR1 and PR2 are of lower order and all three MSE estimators have the same leading term, which is not area-specific. Our results also show that the conditional MSE estimator (8) is much more variable than the model dependent PR MSE estimator, even for moderate area sample sizes.

3.2 Design-based simulations

What happens when, as in real life, we cannot be confident that our data follow a linear mixed model? In order to investigate this situation, we report results from two design-based simulation studies, both based on realistic populations, where a linear model assumption is essentially an approximation. The first involved a sample of 3,591 households spread across $D = 36$ districts of Albania that participated in the 2002 Albanian Living Standards Measurement Study. This sample was bootstrapped to create a realistic population of $N = 724,782$ households by re-sampling with replacement with probability proportional to a household's sample weight. A total of $K = 1,000$ independent stratified random samples were then drawn from this bootstrap population, with total sample size equal to that of the original sample and with districts defining the strata. Sample sizes within districts were the same as in the original sample, and varied between 8 and 688 (with median district sample size equal to 56). The Y variable of interest was household per capita consumption expenditure (HCE) and X was defined by three zero-one variables (ownership of television, parabolic antenna and land). The aim was to estimate the average value of HCE for each district. In the original 2002 survey, the linear relationship between HCE and the three variables making up X was rather weak, with very low predictive power. In particular, only ownership of land was significantly related to HCE at the five percent level. This fit was considerably improved by extending the linear model to include random intercepts, defined by independent district effects. These explained approximately 10 per cent of the residual variation in this model.

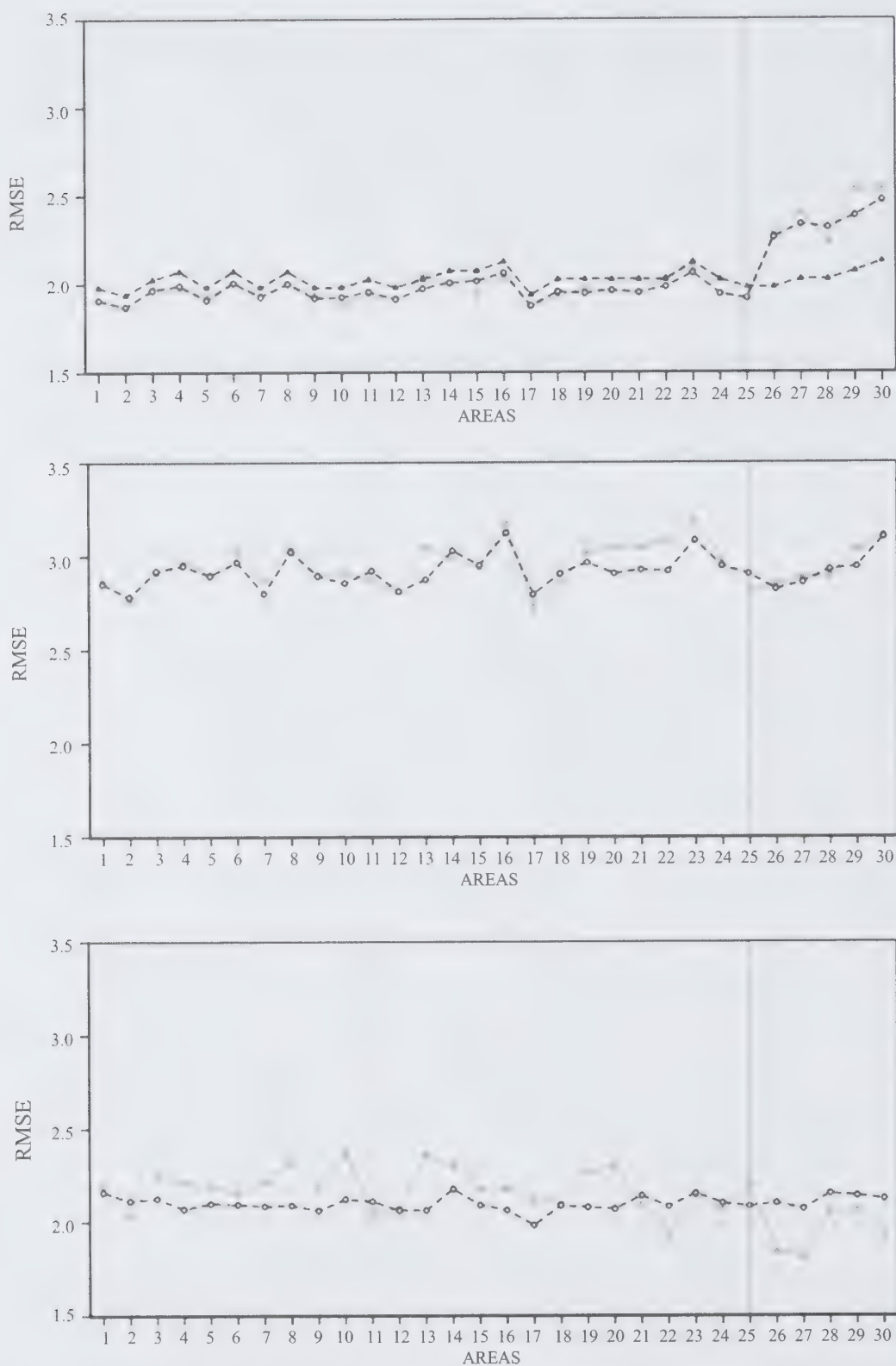


Figure 1 Area specific values of true RMSE (solid line) and average estimated RMSE (dashed line) obtained in the mixture-based simulations SIM3-A and SIM3-A*. Values for the PR0 estimator are indicated by Δ while those for the conditional estimator are indicated by \circ . Plots show results for the EBLUP (top), MBDE (centre) and M-quantile (bottom) estimators. Vertical line separates areas 26-30 with 'outlier' effects from 'well-behaved' areas 1-25. Total sample size is 600 with area-specific sample sizes equal to 20

The second design-based simulation study was based on an ‘outlier free’ version of the population of Australian broadacre farms that was used in the simulation studies reported in Chambers and Tzavidis (2006) and Chandra and Chambers (2009). In particular, this population was defined by bootstrapping a sub-sample of 1,579 ‘non-outlier’ farms that participated in the Australian Agricultural and Grazing Industries Survey (AAGIS) to create a population of $N = 78,072$ farms by re-sampling from the original AAGIS sample with probability proportional to a farm’s sample weight. The small areas of interest in this case were the $D = 28$ broadacre farming regions represented in this sub-sample. The design-based simulation was carried out by selecting $K = 1,000$ independent stratified random samples from this bootstrap population, with strata defined by the regions and with stratum sample sizes defined by those in the original AAGIS sample. These sample sizes vary from 6 to 117, with a median region sample size of 53. Here Y is Total Cash Costs (TCC) associated with operation of the farm, and X is a vector that includes farm area (Area), effects for six post-strata defined by three climatic zones and two farm size bands as well as the interactions of these variables. In the original AAGIS sample the relationship between TCC and Area varies significantly between the six post-strata, with an overall R^2 value of approximately 0.46 after the deletion of two outliers. The fixed effects in the prediction model were therefore specified as corresponding to a separate linear fit of TCC in terms of Area in each post-stratum. Random effects (necessary for computation of the EBLUP and the MBDE, but not the M-quantile predictor) were defined as independent regional effects (*i.e.*, a random intercepts specification) on the basis that in the original AAGIS sample the between region variance component explains about 3 per cent of the total residual variability with the two outliers removed. The aim was to estimate the regional averages of TCC.

Tables 5 and 6 show the median relative biases and the median relative RMSEs of different estimators and corresponding estimators of the MSEs of these estimators based on the $K = 1,000$ independent stratified samples taken from the Albanian and AAGIS populations respectively. It is noteworthy that in spite of the fact that the linear mixed models fitted to both the Albanian and AAGIS data appear reasonable, the gains from adoption of SAE methods based on them do not lead to substantial improvements in efficiency given the original regional sample sizes for these surveys. On the other hand, the M-quantile estimator, which is not based on a random effects specification, works well both in terms of bias and MSE for the AAGIS population in this case (Table 6, Median $n_i = 53$), while the EBLUP, although the best performer in terms of MSE for the Albanian population (Table 5, Median $n_i = 56$), also records the highest biases (albeit still small, with the largest less than 2%) for both populations given the original area sample sizes. The survey regression estimator performs well, although for both populations there are indirect estimators that perform somewhat better. Design-based simulations based on the Albanian and AAGIS populations were also carried out using smaller area sample sizes than in the original surveys. In particular, the overall sample size was reduced for the Albanian population to $n = 291$ (with a median district sample size of 9). Similarly, the overall sample size was reduced for the AAGIS population to $n = 233$ (with a median regional sample size of 8). As expected the RMSE of the point estimators increases as the area sample sizes decrease. Overall, the EBLUP improves its RMSE performance relative to all other estimators given these smaller sample sizes. However, since the realism of these reduced sample size designs is somewhat questionable, we do not place too much emphasis on results derived from them, noting only that they are useful for assessing the performance of MSE estimators with realistic data and with very small sample sizes.

Table 5
Performances of estimators of regional means and their MSE estimators – Albanian household population

Weighting Method	Median $n_i = 56$		Median $n_i = 9$	
	RB(m)	RRMSE(m)	RB(m)	RRMSE(m)
Regression	0.04	6.25	-0.13	16.56
EBLUP, (13)	0.42	5.90	1.62	12.42
MBDE, (18)	0.03	6.14	0.04	16.92
M-quantile, (22)	0.04	6.07	-0.05	16.60
Method/MSE	RB(M)	RRMSE(M)	RB(M)	RRMSE(M)
Regression /VReg	17.6	42	11.2	42
EBLUP/PR0	14.6	44	10.5	50
EBLUP/PR1	14.4	43	8.8	48
EBLUP/PR2	14.5	43	9.7	48
EBLUP/Conditional	0.1	24	7.7	99
MBDE/Conditional	-0.8	25	-5.5	64
M-quantile/Conditional	2.9	27	-2.0	75

Table 6
Performances of estimators of regional means and their MSE estimators – AAGIS farm population

Weighting Method	Median $n_i = 53$		Median $n_i = 8$	
	RB(m)	RRMSE(m)	RB(m)	RRMSE(m)
Regression	0.03	13.36	0.08	29.83
EBLUP, (13)	1.64	13.53	0.92	25.82
MBDE, (18)	-0.73	14.26	-1.02	37.77
M-quantile, (22)	-0.04	11.68	-0.15	32.22
Method/MSE	RB(M)	RRMSE(M)	RB(M)	RRMSE(M)
Regression /VReg	74.1	406	54.7	867
EBLUP/PR0	22.4	131	17.7	374
EBLUP/PR1	19.5	137	19.0	367
EBLUP/PR2	21.0	123	31.1	444
EBLUP/Conditional	5.5	132	17.8	255
MBDE/Conditional	-0.5	181	0.9	318
M-quantile/Conditional	-0.7	69	-1.9	212

Focusing on the simulation results obtained using the original regional sample sizes, we see that all three PR-based MSE estimators for the EBLUP display a substantial upward bias in both sets of design-based simulations as well as larger (Albanian population, Table 5) or comparable (AAGIS population, Table 6) instability to the conditional MSE estimators. For the Albanian population all three versions of the conditional MSE estimator are essentially unbiased whereas for the AAGIS population all three versions of the conditional MSE estimator display small or moderate bias.

It is noteworthy that for the Albanian population (Table 5) the relative performances of the PR MSE estimators improve with smaller samples. However, this is because the conditional MSE estimators then become more unstable. For these very small area samples the conditional MSE estimator is less biased than the PR MSE estimator (7.7% vs. 10.5%) but is also more unstable (RRMSE of conditional MSE estimator is 99% vs. 50% for the PR MSE estimator). This is, however, not the case for the AAGIS population with median $n_i = 8$. In this case, the PR-based MSE estimators perform badly, with the conditional MSE estimators being both less biased and more stable.

The MSE estimator of the regression estimator exhibits moderate or high bias for both populations and all simulation scenarios. For the Albanian population it appears to be competitive to the other MSE estimators in terms of RRMSE but for the AAGIS population it is clearly less stable than the other MSE estimators. Finally, the conditional MSE estimator of the M-quantile estimator performs well with small relative bias and good stability for all simulation scenarios and both populations with the exception of the Albanian population with median $n_i = 9$ where its RRMSE is 75%.

An insight into the reasons for these differences in behaviour can be obtained by examining the area specific RMSE values displayed in Figure 2 for the Albanian

population and in Figure 3 for the AAGIS population. Note that in both cases the sample sizes are those from the original surveys. Thus, in Figure 2 we see that all three conditional MSE estimators track the district-specific design-based RMSEs of their respective estimators exceptionally well. In contrast, the PR0 estimator does not seem to be able to capture between district differences in the design-based RMSE of the EBLUP. In Figure 3 we see that the conditional estimator of the MSE of the M-quantile estimator performs extremely well in all regions, with the corresponding estimator of the MSE of the MBDE also performing well in all regions except one (region 6) where it substantially overestimates the design-based RMSE of this predictor. This region is noteworthy because samples that are unbalanced with respect to Area within the region lead to negative weights under the assumed linear mixed model. The picture becomes more complex when one considers the region-specific RMSE estimation performance of the EBLUP in Figure 3. Here we see that the conditional estimator of the MSE of the EBLUP clearly tracks the region-specific design-based RMSE of this predictor better than the PR0 MSE estimator. With the exception of region 6 (where sample balance is a problem), there seems to be little regional variation in the value of the PR0 estimator of the RMSE of the EBLUP, indicating a serious bias problem.

As noted earlier, it is not uncommon to want to produce an estimate for a small area where there is no sample. In such cases, one has to rely completely on the correctness of the model specification. In Table 7 we illustrate the importance of this assumption by contrasting the estimation and MSE estimation performances of the EBLUP for sampled areas with that of the Synthetic EBLUP for areas where no sample data are available. Two situations are shown. The first is a modification of the model-based SIM1-A simulation with a small average sample size and with five zero sample areas. The second is a similar small sample modification of the design-based simulation based on the

AAGIS population, with four zero sample areas. It is clear that when the model underpinning the EBLUP actually holds (*i.e.*, SIM1-A), estimation and MSE estimation (either based on PR0, or on the conditional alternative) works well. The problem is that when there is some doubt about how well this model holds (as in the AAGIS population), then the EBLUP can fail, and our estimator of its MSE can also

fail to identify this problem. This is nicely illustrated by the results for the AAGIS population in Table 7 where we see that both the PR0 and conditional MSE estimators for the Synthetic EBLUP completely fail to identify the large positive bias of the Synthetic EBLUP and so end up with a large downward bias.

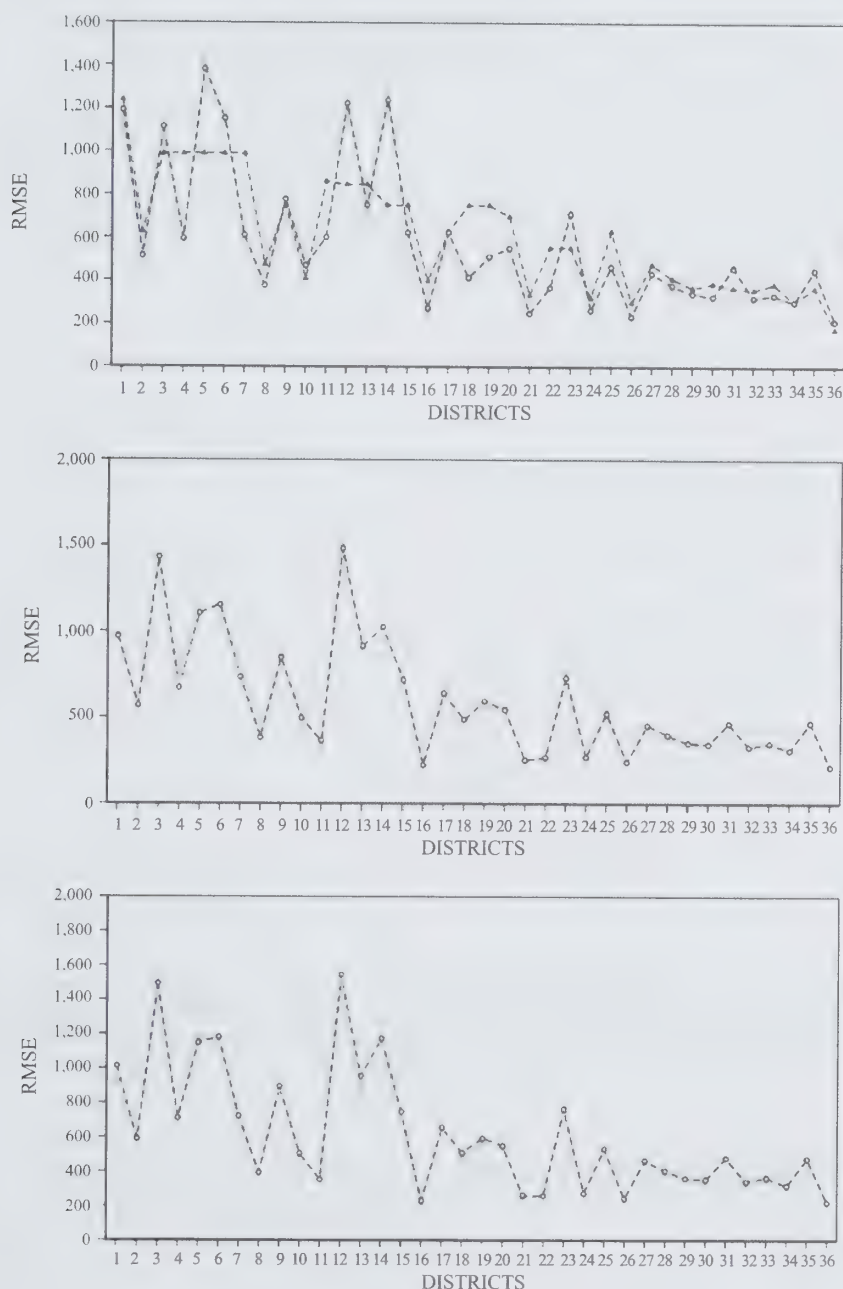


Figure 2 District level values of true design-based RMSE (solid line) and average estimated RMSE (dashed line) obtained in the design-based simulations using the Albanian household population. Districts are ordered in terms of increasing population size. Values for the PR0 estimator are indicated by Δ while those for the conditional estimator are indicated by \circ . Plots show results for the EBLUP (top), MBDE (centre) and M-quantile (bottom) estimators

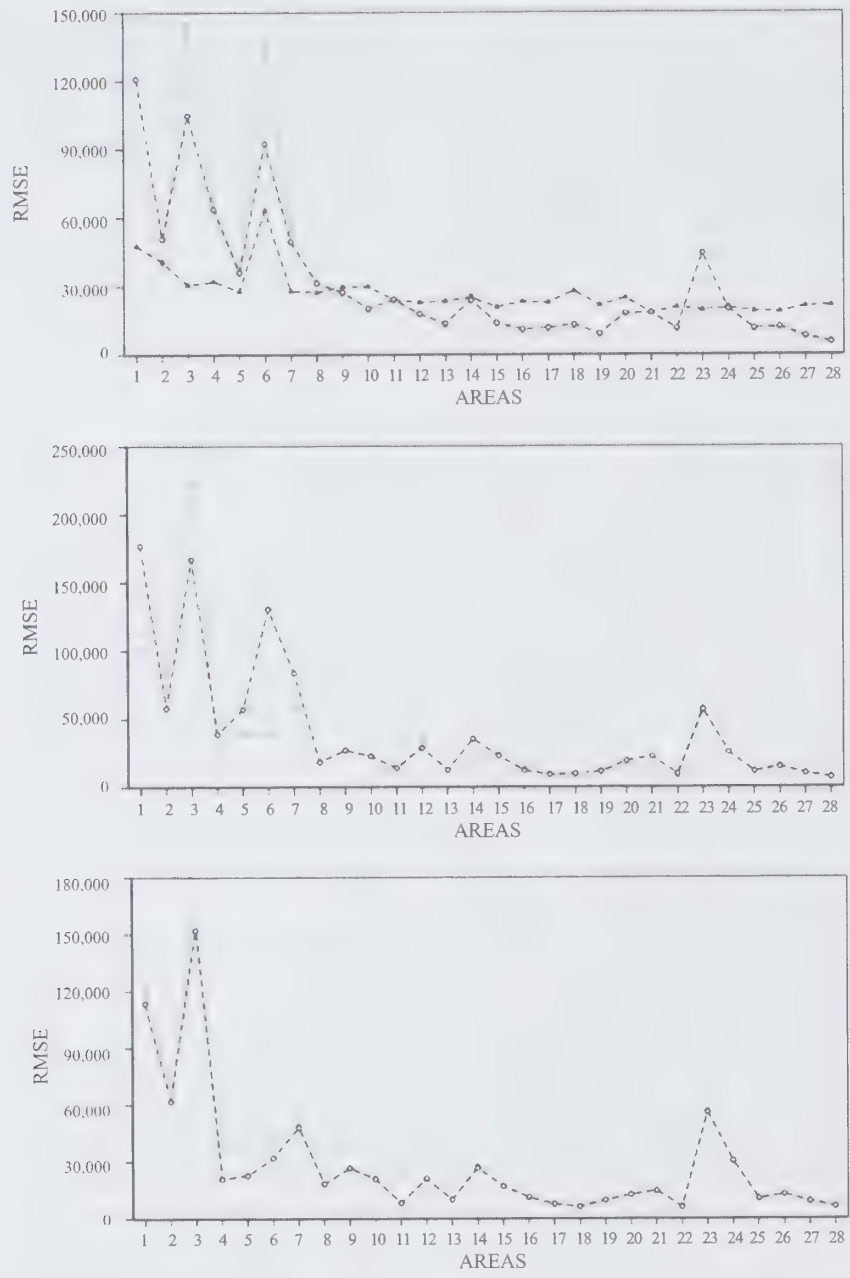


Figure 3 Regional values of true design-based RMSE (solid line) and average estimated RMSE (dashed line) obtained in the design-based simulations using the AAGIS farm population. Regions are ordered in terms of increasing population size. Values for the PR0 estimator are indicated by Δ while those for the conditional estimator are indicated by \circ . Plots show results for the EBLUP (top), MBDE (centre) and M-quantile (bottom) estimators

Table 7 Performance of EBLUP and MSE estimators when there are areas with zero sample

Weighting Method/ Estimator		SIM1-A, median $n_i = 10$		AAGIS, median $n_i = 9$	
		RB(m)	RRMSE(m)	RB(m)	RRMSE(m)
Areas with $n_i > 0$	(13)/EBLUP	0.00	0.52	2.29	24.94
Areas with $n_i = 0$	(23)/Synthetic EBLUP	-0.05	1.25	87.45	96.46
MSE Estimator		RB(M)	RRMSE(M)	RB(M)	RRMSE(M)
Areas with $n_i > 0$	(13)/PR0	0.5	11	29.91	760
	(13)/Conditional	0.7	50	23.87	298
Areas with $n_i = 0$	(23)/PR0	-1.8	35	-29.07	601
	(23)/Conditional	-3.6	34	-31.45	101

4. Conclusions and discussion

In this paper we propose a bias-robust and easily implemented method of estimating the conditional MSE of pseudo-linear estimators of small area means (and totals). Our empirical results show that this method of MSE estimation performs reasonably well in terms of bias when used to estimate the model-based MSE and the design-based MSE of the three rather different pseudo-linear estimators considered in this paper. However, this improved bias performance comes at the cost of increased variability. In particular, when area sample sizes are very small, we do not recommend use of our proposed method of MSE estimation for a conditionally biased estimator like the EBLUP.

The EBLUP is widely used in SAE, and in this context the model-dependent MSE estimator PR0 for the EBLUP suggested by Prasad and Rao (1990) is unbiased when its model assumptions are valid (SIM1-A/B and SIM2-A/B in our model-based simulations) but is biased in the presence of outlier area effects (SIM3-A/A* and SIM3-B/B*). It was also the most stable MSE estimator in the model-based simulations. However, its area-averaged construction meant that it did not track the area-specific MSE of the EBLUP in both our design-based simulations, where the correctness of the assumed linear mixed model could only be considered as approximate. This suggests that our proposed conditional MSE estimation method should be considered as an alternative to PR0 in situations where there is some doubt about the correctness of the specification of the small area linear mixed model or where the area sample sizes are not small. Some idea of what constitutes a small sample size can be deduced from the empirical results presented in this paper.

If there is doubt about the validity of the assumed linear mixed model, the user could consider estimation based on a more widely applicable alternative model, *e.g.*, the M-quantile model, or replace the EBLUP by a more outlier-robust alternative (Sinha and Rao 2009). In the former case the approach that we propose in this paper is currently the only analytical approach to MSE estimation, while in the latter case it provides an analytic alternative to more computationally intensive bootstrap methods of MSE estimation. Note however, that for very small area-specific sample sizes the bias-robust MSE estimator proposed in this paper remains unstable.

A future line of research could be to compare the analytic MSE estimation method proposed in this paper with bootstrap-based MSE estimators, *e.g.*, the nonparametric bootstrap MSE estimator of the M-quantile estimator proposed by Tzavidis, Marchetti and Chambers (2010), and the bootstrap MSE estimator for the Robust EBLUP estimator proposed by Sinha and Rao (2009). A key issue in

this investigation will be to investigate whether alternative bootstrap MSE estimators are more stable, especially for small area-specific sample sizes.

The extension of the conditional MSE approach to non-linear SAE situations remains to be done. However, since this approach is closely linked to robust population level MSE estimation based on Taylor series linearisation (as well as jackknife estimation of MSE, see Valliant, Dorfman and Royall 2000, section 5.4.2), it should be possible to develop appropriate extensions for corresponding small area non-linear estimation methods. Although the relevant results are not provided here, some evidence for this is that the conditional MSE estimation method described in this paper has already been used to estimate the MSE of the MBDE when it is applied to variables that do not lend themselves to linear mixed modelling, *e.g.*, those with a high proportion of zero values (Chandra and Chambers 2009), and categorical variables (Chandra, Chambers and Salvati 2011). More recently, the approach has also been used to estimate the MSE of geographically weighted M-quantile small area estimators in situations where the small area values are spatially correlated (Salvati, Tzavidis, Pratesi and Chambers 2011). It has also been used by Salvati, Chandra, Ranalli and Chambers (2010) to estimate the MSE of small area estimators based on a nonparametric small area model (Opsomer, Claeskens, Ranalli, Kauermann and Breidt 2008).

As is clear from the development in this paper, our preferred approach to MSE estimation assumes that the MSE of real interest is that defined by the area-specific model (1). This is in contrast to the usual approach to defining MSE in SAE, which adopts an area-averaged MSE concept as the appropriate measure of the accuracy of a small area estimator. As pointed out by Longford (2007), the ultimate aim in SAE is to make inferences about small area characteristics conditional on the realised (but unknown) values of small area effects, *i.e.*, with respect to (1). One can consider this to be a design-based objective (as in Longford 2007), or, as we prefer, a model-based objective that does not quite fit into the usual random effects framework for SAE. In either case we are interested in variability that is with respect to fixed area-specific expected values. This is consistent with the concept of variability that is typically applied in population level inference.

Acknowledgements

The authors would like to acknowledge the valuable comments and suggestions of the Editor, the Associate Editor and two referees. These led to a considerable improvement in the paper.

References

- Chambers, R.L., and Dunstan, R. (1986). Estimating distribution functions from survey data. *Biometrika*, 73, 597-604.
- Chambers, R., and Tzavidis, N. (2006). M-quantile models for small area estimation. *Biometrika*, 93, 255-268.
- Chandra, H., and Chambers, R. (2009). Multipurpose weighting for small area estimation. *Journal of Official Statistics*, 25(3), 379-395.
- Chandra, H., Chambers, R. and Salvati, N. (2011). Small area estimation of proportions in business surveys. To appear in *Journal of Statistical Computation and Simulation*.
- Henderson, C.R. (1953). Estimation of variance and covariance components. *Biometrics*, 9, 226-252.
- Longford, N.T. (2007). On standard errors of model-based small-area estimators. *Survey Methodology*, 33, 69-79.
- Opsomer, J.D., Claeskens, G., Ranalli, M.G., Kauermann, G. and Breidt, F.J. (2008). Nonparametric small area estimation using penalized spline regression. *Journal of the Royal Statistical Society, Series B*, 70, 265-286.
- Prasad, N.G.N., and Rao, J.N.K. (1990). The estimation of the mean squared error of small area estimators. *Journal of the American Statistical Association*, 85, 163-171.
- Rao, J.N.K. (2003). *Small Area Estimation*. New York: John Wiley & Sons, Inc.
- Royall, R.M. (1976). The linear least squares prediction approach to two-stage sampling. *Journal of the American Statistical Association*, 71, 657-664.
- Royall, R.M., and Cumberland, W.G. (1978). Variance estimation in finite population sampling. *Journal of the American Statistical Association*, 73, 351-358.
- Salvati, N., Chandra, H., Ranalli, M.G. and Chambers, R. (2010). Small area estimation using a nonparametric model based direct estimator. *Computational Statistics and Data Analysis*, 54, 2159-2171.
- Salvati, N., Tzavidis, N., Pratesi, M. and Chambers, R. (2011). Small area estimation via M-quantile geographically weighted regression. Forthcoming in *TEST*, DOI 10.1007/s11749-010-0231-1.
- Sinha, S.K., and Rao, J.N.K. (2009). Robust small area estimation. *The Canadian Journal of Statistics*, 37(3), 381-399.
- Street, J.O, Carroll, R.J. and Ruppert, D. (1988). A note on computing robust regression estimates via iteratively reweighted least squares. *The American Statistician*, 42, 152-154.
- Tzavidis, N., Marchetti, S. and Chambers, R. (2010). Robust prediction of small area means and distributions. *Australian and New Zealand Journal of Statistics*, 52, 167-186.
- Valliant, R., Dorfman, A.H. and Royall, R.M. (2000). *Finite Population Sampling and Inference*. New York: John Wiley & Sons, Inc.

Variance estimation under composite imputation: The methodology behind SEVANI

Jean-François Beaumont and Joël Bissonnette¹

Abstract

Composite imputation is often used in business surveys. The term “composite” means that more than a single imputation method is used to impute missing values for a variable of interest. The literature on variance estimation in the presence of composite imputation is rather limited. To deal with this problem, we consider an extension of the methodology developed by Särndal (1992). Our extension is quite general and easy to implement provided that linear imputation methods are used to fill in the missing values. This class of imputation methods contains linear regression imputation, donor imputation and auxiliary value imputation, sometimes called cold-deck or substitution imputation. It thus covers the most common methods used by national statistical agencies for the imputation of missing values. Our methodology has been implemented in the System for the Estimation of Variance due to Nonresponse and Imputation (SEVANI) developed at Statistics Canada. Its performance is evaluated in a simulation study.

Key Words: Auxiliary value imputation; Composite imputation; Donor imputation; Imputation model; Linear imputation; Regression imputation; SEVANI.

1. Introduction

Composite imputation is often used in business surveys. The term “composite” means that more than a single imputation method is used to impute missing values for a variable of interest. The choice of a method over another one depends on the availability of auxiliary variables. For instance, ratio imputation could be used to impute a missing value when an auxiliary value is available; otherwise, mean imputation could be an alternative.

The problem of estimating the variance in the presence of a single imputation method has been extensively studied in the literature; *e.g.*, two excellent reviews of this topic are: Lee, Rancourt and Särndal (2001) and Haziza (2009). Although the use of composite imputation occurs frequently in practice, there is little literature on estimating its variance. The literature includes a jackknife variance estimator that was proposed and evaluated empirically in Rancourt, Lee and Särndal (1993). Sitter and Rao (1997) developed further the theory and obtained design-consistent linearization and jackknife variance estimators. In both papers, two imputation methods were considered, with ratio imputation being one of the methods, simple random sampling was used and uniform nonresponse was assumed. Later, Felx and Rancourt (2001) extended the general methodology proposed in Särndal (1992) and Deville and Särndal (1994) to composite imputation using simplifying assumptions. Finally, Shao and Steel (1999) developed an interesting and general reverse approach to variance estimation to deal with composite imputation (see also Kim and Rao 2009). Shao and Steel (1999) claimed that their reverse approach leads to

derivations that are less involved than those found in Deville and Särndal (1994). We do not fully agree with this statement. Our results indicate that, in general, our extension to Särndal’s approach actually leads to simpler derivations than those obtained with the Shao and Steel approach. The reverse approach may however become quite attractive when the sampling fraction is negligible and a replication variance estimation technique is chosen (see section 7 for greater detail).

We consider the methodology proposed by Särndal (1992) as a starting point. It requires the validity of an imputation model; *i.e.*, a model for the variable being imputed. At first glance, the extension of this methodology to composite imputation seems to be quite tedious, as noted by Shao and Steel (1999), until we notice that most imputation methods used in practice lead to imputed estimators that are linear in the observed values of the variable of interest. This considerably simplifies the derivation of a variance estimator even when there is a single imputation method. For the estimation of the sampling portion of the overall variance, we use a methodology (see Beaumont and Bocci 2009) that is slightly different than the one proposed by Särndal (1992). This allows us to simplify the derivations further. This research has been implemented in version 2 of the System for the Estimation of Variance due to Nonresponse and Imputation (SEVANI), which is developed at Statistics Canada (see Beaumont, Bissonnette and Bocci 2010).

The paper is structured as follows. In section 2, some notation is introduced and composite imputation is explained. Linear imputation is defined in section 3. Our

1. Jean-François Beaumont, Statistics Canada, Statistical Research and Innovation Division, Tunney’s Pasture, Ottawa, Ontario, Canada, K1A 0T6. E-mail: jean-francois.beaumont@statcan.gc.ca; Joël Bissonnette, Statistics Canada, Business Survey Methods Division, Tunney’s Pasture, Ottawa, Ontario, Canada, K1A 0T6. E-mail: joel.bissonnette@statcan.gc.ca.

approach to inference and our main assumptions are described in section 4. In section 5, a number of results are stated regarding variance estimation under composite imputation. Section 6 presents the results of a simulation study that assesses the performance of our variance estimator. The reverse approach is briefly discussed in section 7 to highlight the differences with our approach. Finally, a short conclusion is given in section 8.

2. What is composite imputation?

Suppose that we are interested in estimating the population domain total $\theta = \sum_{k \in U} d_k y_k$, where U is the finite population of size N , y is the variable of interest and d is a domain indicator variable indicating whether population unit k is in the domain of interest ($d_k = 1$) or not ($d_k = 0$). A sample s of size n is selected from the finite population U according to a probability sampling design $p(s)$. In the absence of missing values, θ can be estimated by the Horvitz-Thompson estimator $\hat{\theta} = \sum_{k \in s} w_k d_k y_k$, where $w_k = 1/\pi_k$ is the design weight and π_k is the selection probability of unit k . Although it is possible to extend our results to calibration estimators, it is not considered in this paper to keep matters simple.

Variable y can be missing for some of the sampled units but we assume that the domain indicator variable d is always observed for those units. The set of sampled units with an observed y -value, called the set of respondents, is denoted by s_r . It is assumed to have been generated according to a nonresponse mechanism $q(s_r | s)$. The set of nonrespondents is denoted by $s_m = s - s_r$. It is further split into J mutually exclusive subsets, $s_m^{(j)}$, $j = 1, \dots, J$, such that $s_m = \bigcup_{j=1}^J s_m^{(j)}$, if composite imputation with $J > 1$ imputation methods is used. All the missing y -values within a given subset $s_m^{(j)}$ are imputed with the same method j . However, different imputation methods are used to impute missing values in different subsets. The resulting imputed estimator can be expressed as

$$\begin{aligned} \hat{\theta}_I &= \sum_{k \in s_r} w_k d_k y_k + \sum_{k \in s_m} w_k d_k y_k^* \\ &= \sum_{k \in s_r} w_k d_k y_k + \sum_{j=1}^J \sum_{k \in s_m^{(j)}} w_k d_k y_k^*, \end{aligned} \quad (2.1)$$

where y_k^* is the imputed y -value for unit k .

Composite imputation is quite frequent in business surveys. It is used because there are missing values in auxiliary variables used for imputation. To fix ideas, let \mathbf{x}_k be the complete vector of auxiliary variables for unit k . Ideally, all the missing y -values would be imputed using a single imputation method based on the complete vector \mathbf{x}_k . Unfortunately, there may be missing values in the auxiliary

variables so that, for some nonrespondents, we cannot use \mathbf{x}_k to impute their missing y -value; we can only use a subset of \mathbf{x}_k . We denote as $\mathbf{x}_k^{\text{obs}}$, the vector of observed auxiliary variables for unit k . This vector does not necessarily contain the same observed variables from one unit to the next. To impute the missing y -value of a given unit k , an imputation method is chosen based on the available auxiliary variables $\mathbf{x}_k^{\text{obs}}$. Since there may be a number of nonresponse patterns in the complete vector of auxiliary variables, the imputation strategy may contain a number of imputation methods.

Example:

The variance estimation issues raised by composite imputation can be better understood by considering the following example. Suppose that the complete vector of auxiliary variables for unit k is $\mathbf{x}_k = (x_{1k}, x_{2k})$, where x_{1k} is strongly related to y_k but subject to missing values while x_{2k} is set to a constant for all sampled units ($x_{2k} = 1, k \in s$). Ideally, x_{1k} is used to impute y_k if it is missing. If x_{1k} is not available, only x_{2k} can be used. Table 1 summarizes the information available for the different subsets of the sample s .

Table 1
Available information when there is one auxiliary variable x_1 and a constant x_2

Subsets		y	x_1	x_2	\mathbf{x}^{obs}
s_r	$s_r^{(1)}$	O	O	O	(x_1, x_2)
	$s_r^{(2)}$	O	M	O	(M, x_2)
s_m	$s_m^{(1)}$	M	O	O	(x_1, x_2)
	$s_m^{(2)}$	M	M	O	(M, x_2)

O: Observed; M: Missing.

The set of nonrespondents s_m is divided into the subsets $s_m^{(1)}$ and $s_m^{(2)}$ depending on the availability of x_1 . Similarly, the set of respondents is divided into subsets $s_r^{(1)}$ and $s_r^{(2)}$. In this example, we could use ratio imputation to impute missing y -values in $s_m^{(1)}$ and mean imputation to impute missing y -values in $s_m^{(2)}$. Note that simple linear regression imputation could be used instead of ratio imputation (if it better fits the data). We have chosen ratio imputation in this example for its simplicity and because it is frequently used in business surveys.

Only the respondents in $s_r^{(1)}$ can be used to impute missing y -values in $s_m^{(1)}$ through ratio imputation. The imputed value for a unit k in $s_m^{(1)}$ is $y_k^* = x_{1k} \sum_{l \in s_r^{(1)}} \omega_l^{(1)} y_l / \sum_{l \in s_r^{(1)}} \omega_l^{(1)} x_{1l}$, where $\omega_l^{(1)}$ is some weight used for ratio imputation (imputation method 1). Typical choices are: $\omega_l^{(1)} = w_l$ (design-weighted imputation) or $\omega_l^{(1)} = 1$ (unweighted imputation). For mean imputation, the respondents in $s_r^{(2)}$ as well as those in $s_r^{(1)}$ can be used to impute

missing y -values in $s_m^{(2)}$. In practice, it is common to use both sets of respondents to improve the stability of the imputed mean. The imputed value for a unit k in $s_m^{(2)}$ is

$$y_k^* = \sum_{l \in s_r} \omega_l^{(2)} y_l / \sum_{l \in s_r} \omega_l^{(2)},$$

where $\omega_l^{(2)}$ is a weight used for mean imputation (imputation method 2). (Typical choices of $\omega_l^{(2)}$ are the same as those for $\omega_l^{(1)}$; i.e., $\omega_l^{(2)} = w_l$ or $\omega_l^{(2)} = 1$.) This implies that units in $s_r^{(1)}$ can be contributors to both imputation methods. This raises issues for variance estimation of the resulting composite imputation estimator. These issues will be addressed in section 5.

3. What is linear imputation?

The imputation method j is said to be linear if the imputed value y_k^* for a sample unit $k \in s_m^{(j)}$ can be written in the linear form

$$y_k^* = \phi_{0k}^{(j)} + \sum_{l \in s_r} \phi_{lk}^{(j)} y_l. \quad (3.1)$$

The quantities $\phi_{0k}^{(j)}$ and $\phi_{lk}^{(j)}$, for $l \in s_r$, are obtained without using y -values, but may depend on s and s_r . The linear form (3.1) is satisfied by several of the most common imputation methods in practice such as (weighted or unweighted) linear regression imputation, donor imputation and auxiliary value imputation. A nice review of these methods is found in Haziza (2009). Note that auxiliary value imputation does not use the y -values of respondents; i.e., $y_k^* = \phi_{0k}^{(j)}$ (see Beaumont, Haziza and Bocci 2011). For donor imputation, the imputed value y_k^* is equal to the y -value of a suitably chosen respondent (donor) so that $\phi_{0k}^{(j)} = 0$ and $\phi_{lk}^{(j)} = 0$ for all but one respondent $l \in s_r$. Detailed expressions for $\phi_{0k}^{(j)}$ and $\phi_{lk}^{(j)}$ are given in the Methodology Guide of SEVANI (Beaumont, Bissonnette and Bocci 2010), which is available on request from the authors.

Let $\Omega_I^{(j)} = \sum_{k \in s_m^{(j)}} w_k d_k y_k^*$ be the contribution of imputation method j to the estimator $\hat{\theta}_I$. Using (3.1), $\Omega_I^{(j)}$ can be decomposed as follows:

$$\begin{aligned} \Omega_I^{(j)} &= \sum_{k \in s_m^{(j)}} w_k d_k y_k^* \\ &= \sum_{k \in s_m^{(j)}} w_k d_k \phi_{0k}^{(j)} + \sum_{l \in s_r} y_l \sum_{k \in s_m^{(j)}} w_k d_k \phi_{lk}^{(j)} \\ &= W_{0d}^{(j)} + \sum_{l \in s_r} W_{dl}^{(j)} y_l, \end{aligned} \quad (3.2)$$

where $W_{0d}^{(j)} = \sum_{k \in s_m^{(j)}} w_k d_k \phi_{0k}^{(j)}$ and $W_{dl}^{(j)} = \sum_{k \in s_m^{(j)}} w_k d_k \phi_{lk}^{(j)}$. Using (3.2), the imputed estimator (2.1) can be expressed in the linear form:

$$\begin{aligned} \hat{\theta}_I &= \sum_{k \in s_r} w_k d_k y_k + \sum_{j=1}^J \Omega_I^{(j)} \\ &= W_{0d}^{(+)} + \sum_{k \in s_r} (w_k d_k + W_{dk}^{(+)}) y_k, \end{aligned} \quad (3.3)$$

where $W_{0d}^{(+)} = \sum_{j=1}^J W_{0d}^{(j)}$ and $W_{dk}^{(+)} = \sum_{j=1}^J W_{dk}^{(j)}$.

Continuing with the example introduced at the end of section 2, we observe that, for ratio imputation, $\phi_{0k}^{(1)} = 0$ and $\phi_{lk}^{(1)} = \omega_l^{(1)} x_{lk} / \sum_{l \in s_r} \omega_l^{(1)} x_{lk}$, for $l \in s_r$, with $\omega_l^{(1)} = 0$, for $l \in s_r^{(2)}$. For mean imputation, we have $\phi_{0k}^{(2)} = 0$ and $\phi_{lk}^{(2)} = \omega_l^{(2)} / \sum_{l \in s_r} \omega_l^{(2)}$, for $l \in s_r$. Consequently, $W_{0d}^{(1)} = 0$, $W_{0d}^{(2)} = 0$,

$$W_{dl}^{(1)} = \omega_l^{(1)} \sum_{k \in s_m^{(1)}} w_k d_k x_{lk} / \sum_{k \in s_r} \omega_k^{(1)} x_{lk}$$

and $W_{dl}^{(2)} = \omega_l^{(2)} \sum_{k \in s_m^{(2)}} w_k d_k / \sum_{k \in s_r} \omega_k^{(2)}$. This implies that $W_{0d}^{(+)} = 0$ and $W_{dk}^{(+)} = W_{dk}^{(1)} + W_{dk}^{(2)}$.

4. Approach to inference and main assumptions

We consider three sources of variability when evaluating expectations and variances of the imputed estimator: the variability due to the imputation model, the sampling design and the nonresponse mechanism. Note that the use of an imputation model to make inference in the presence of imputation can be found in Rubin (1987), Hidirolou (1989) and Särndal (1992). In what follows, we will use the subscripts m , p and q to denote the expectations, variances and covariances evaluated with respect to the imputation model, sampling design and nonresponse mechanism respectively.

We consider the following imputation model to describe the relationship between the y -variable and the vector \mathbf{x}^{obs} of observed auxiliary variables:

$$\begin{aligned} E_m(y_k | \mathbf{X}^{\text{obs}}) &= \mu_k \\ V_m(y_k | \mathbf{X}^{\text{obs}}) &= \sigma_k^2 \\ \text{cov}_m(y_k, y_l | \mathbf{X}^{\text{obs}}) &= 0, \end{aligned} \quad (4.1)$$

for $k \neq l$ and $k, l \in U$. The population matrix \mathbf{X}^{obs} contains the vectors of observed auxiliary variables, $\mathbf{x}_k^{\text{obs}}$, for $k \in U$, and μ_k and σ_k^2 are functions of $\mathbf{x}_k^{\text{obs}}$. Asymptotically m -unbiased and m -consistent estimators of μ_k and σ_k^2 are denoted by $\hat{\mu}_k$ and $\hat{\sigma}_k^2$ respectively. Since we will always condition on \mathbf{X}^{obs} , we exclude this conditioning from the notation to simplify it. For instance, $E_m(y_k | \mathbf{X}^{\text{obs}})$ will be written as $E_m(y_k)$.

In model (4.1), we condition on the observed auxiliary variables. Since the nonresponse pattern in the vector \mathbf{x} is not the same for all the nonrespondents, a separate conditional model must be validated and fitted for each nonresponse pattern. In principle, these conditional models should be used to determine the imputation methods chosen.

Note that model (4.1) reduces to the standard conditional model (e.g., Särndal 1992) when the vector \mathbf{x} of auxiliary variables is not subject to missing values.

Remark: The validity of the variance estimation method in section 5 requires μ_k and σ_k^2 to be correctly specified. Although a parametric form for μ_k may often be acceptable, it may be more difficult to determine a suitable parametric form for σ_k^2 . To avoid this issue and obtain some robustness against misspecification of the model variance, σ_k^2 can be estimated non parametrically; see the empirical study of Beaumont, Haziza and Bocci (2011) for an illustration of this property under auxiliary value imputation. In the context of donor imputation, Beaumont and Bocci (2009) showed empirically that nonparametric estimation of both μ_k and σ_k^2 , via penalized smoothing splines, reduced significantly the vulnerability of our variance estimator to misspecifications of the model mean and variance.

In addition to the imputation model (4.1), we also assume that:

$$F(\mathbf{Y} \mid s, s_r, \mathbf{X}^{\text{obs}}, \mathbf{Z}, \mathbf{D}) = F(\mathbf{Y} \mid \mathbf{X}^{\text{obs}}), \quad (4.2)$$

where $F(\cdot)$ denotes the distribution function, \mathbf{Y} and \mathbf{D} are N -element vectors containing respectively y_k and d_k as their k^{th} element, and \mathbf{Z} is a N -row matrix of design information, which implicitly or explicitly contains information about the selection probabilities π_k and joint selection probabilities π_{kl} , for $k, l \in U$. This assumption, often implicit in other papers, allows us to treat the response indicators, the domain indicators and the design information as fixed when taking model expectations and variances. A careful choice of the auxiliary variables is necessary to satisfy this assumption. For instance, the design information and the domain indicators should be considered as potential auxiliary variables.

The imputation strategy given in our example started in section 2 could be justified by a model with $\mu_k = \beta_1 x_{1k}$ and $\sigma_k^2 = \sigma_1^2 x_{1k}$, for $k \in s_r^{(1)}$ or $k \in s_m^{(1)}$, and $\mu_k = \beta_2$ and $\sigma_k^2 = \sigma_2^2$, for $k \in s_r^{(2)}$ or $k \in s_m^{(2)}$. The model parameters β_1 , β_2 , σ_1^2 and σ_2^2 are unknown. Note that if the x_{1k} 's are assumed to be identically distributed random variables with mean μ_x and variance σ_x^2 , then $\beta_2 = \beta_1 \mu_x$ and $\sigma_2^2 = \beta_1^2 \sigma_x^2 + \sigma_1^2 \mu_x$. The imputed values $y_k^* = \hat{\mu}_k$, for $k \in s_m$, are obtained by estimating the model parameters β_1 and β_2 from the observed data. For instance, the m -unbiased estimators of β_1 and β_2 could be chosen as

$$\hat{\beta}_1 = \sum_{k \in s_r^{(1)}} \omega_k^{(1)} y_k / \sum_{k \in s_r^{(1)}} \omega_k^{(1)} x_{1k}$$

and

$$\hat{\beta}_2 = \sum_{k \in s_r^{(2)}} \omega_k^{(2)} y_k / \sum_{k \in s_r^{(2)}} \omega_k^{(2)}$$

respectively. This would lead to $\hat{\mu}_k = \hat{\beta}_1 x_{1k}$, for $k \in s_r^{(1)}$ or $k \in s_m^{(1)}$, and $\hat{\mu}_k = \hat{\beta}_2$, for $k \in s_r^{(2)}$ or $k \in s_m^{(2)}$. As in section 2, one could also consider the potentially more efficient estimator $\hat{\beta}_2^* = \sum_{k \in s_r} \omega_k^{(2)} y_k / \sum_{k \in s_r} \omega_k^{(2)}$ instead of $\hat{\beta}_2$. Unfortunately, $\hat{\beta}_2^*$ is biased under the model since

$$E_m(\hat{\beta}_2^* \mid s, s_r) = \beta_2 + \frac{\sum_{k \in s_r^{(1)}} \omega_k^{(2)} (x_{1k} \beta_1 - \beta_2)}{\sum_{k \in s_r} \omega_k^{(2)}}. \quad (4.3)$$

As pointed out above, if the x_{1k} 's are assumed to be identically distributed random variables with mean μ_x and variance σ_x^2 , $\beta_2 = \beta_1 \mu_x$ and equation (4.3) can be rewritten as

$$E_m(\hat{\beta}_2^* \mid s, s_r) = \beta_2 + \beta_1 \frac{\sum_{k \in s_r^{(1)}} \omega_k^{(2)} \sum_{k \in s_r^{(1)}} \omega_k^{(2)} (x_{1k} - \mu_x)}{\sum_{k \in s_r} \omega_k^{(2)} \sum_{k \in s_r^{(1)}} \omega_k^{(2)}}. \quad (4.4)$$

It can be shown under weak conditions that $E_m(\hat{\beta}_2^* \mid s, s_r) = \beta_2 + O_p(1/\sqrt{n})$ so that the model bias of $\hat{\beta}_2^*$ is asymptotically negligible. However, since $\text{var}_m(\hat{\beta}_2^* \mid s, s_r) = O_p(1/n)$, the squared model bias is not necessarily asymptotically negligible compared to the model variance of $\hat{\beta}_2^*$. At least, $\hat{\beta}_2^*$ is m -consistent for β_2 . From (4.3) or (4.4), we can see that the model bias of $\hat{\beta}_2^*$ can be controlled by assigning a smaller weight $\omega_k^{(2)}$ to units $k \in s_r^{(1)}$ relative to units $k \in s_r^{(2)}$. For instance, one could consider using $\omega_k^{(2)} = w_k / n^\alpha$, for $k \in s_r^{(1)}$ and some $\alpha > 0$, and $\omega_k^{(2)} = w_k$, for $k \in s_r^{(2)}$. In the extreme case where $\omega_k^{(2)} = 0$, for $k \in s_r^{(1)}$, $\hat{\beta}_2^*$ is model-unbiased because it is equal to $\hat{\beta}_2$. Note that the model bias of $\hat{\beta}_2^*$ could be larger than $O_p(1/\sqrt{n})$ if x_{1k} , $k \in s_r^{(1)}$, have a mean different from μ_x , $k \in s_r^{(2)}$. In such case, controlling the model bias of $\hat{\beta}_2^*$ might be more important.

In the case of donor imputation, a fourth source of variability needs to be considered when donors are randomly selected among respondents to impute nonrespondents. In this paper, the subscript q will implicitly indicate that moments are evaluated with respect to the joint distribution induced by the nonresponse mechanism and the random donor selection mechanism. As a result, when conditioning on s_r , as in (4.2), it should be kept in mind that conditioning is not only on the set of respondents but also on the set of selected donors.

5. Variance estimation

Särndal (1992) expresses the total error of the imputed estimator as:

$$\hat{\theta}_I - \theta = (\hat{\theta} - \theta) + (\hat{\theta}_I - \hat{\theta}), \quad (5.1)$$

where the first term on the right-hand side of (5.1) is called the sampling error and the second term is called the nonresponse error. Using the assumptions given in section 4 and $E_p(\hat{\theta} - \theta) = 0$, the overall bias of the imputed estimator reduces to $E_{mpq}(\hat{\theta}_I - \theta) = E_{pq} B_m$, where $B_m = E_m(\hat{\theta}_I - \hat{\theta} | s, s_r)$ is the (conditional) model bias of the imputed estimator. Using (2.1), the model bias can be expressed as

$$B_m = \sum_{j=1}^J \sum_{k \in s_m^{(j)}} w_k d_k E_m(y_k^* - y_k | s, s_r). \quad (5.2)$$

This means that the model bias and the overall bias vanish if the model expectation of the imputation error, $y_k^* - y_k$, is zero, for $k \in s_m^{(j)}$ and $j = 1, \dots, J$. In principle, an imputation strategy should be chosen so that this condition is satisfied (at least approximately). This is typically assumed in the literature (e.g., Särndal 1992; Shao and Steel 1999).

In the example introduced in section 2, the model bias (5.2) reduces to

$$B_m = \left(\sum_{k \in s_m^{(2)}} w_k d_k \right) E_m(\hat{\beta}_2^* - \beta_2 | s, s_r).$$

An expression for $E_m(\hat{\beta}_2^* - \beta_2 | s, s_r)$ is given by (4.3) or (4.4). As noted in the paragraph that follows equation (4.4), the model bias, B_m , can be controlled by assigning a smaller weight $\omega_k^{(2)}$ to units $k \in s_r^{(1)}$ relative to units $k \in s_r^{(2)}$. It is also small if the number of nonrespondents imputed by method 2 is small. Note that our variance (or Mean Squared Error, MSE) estimation approach requires the slightly weaker assumption that $E_q(B_m | s)$ is negligible (see section 5.3).

Using (5.1), Särndal (1992) decomposed the overall MSE into three components:

$$E_{mpq}(\hat{\theta}_I - \theta)^2 = E_m \text{var}_p(\hat{\theta}) + E_{pq} E_m \{(\hat{\theta}_I - \hat{\theta})^2 | s, s_r\} + 2 E_{pq} E_m \{(\hat{\theta}_I - \hat{\theta})(\hat{\theta} - \theta) | s, s_r\}. \quad (5.3)$$

The overall MSE (5.3) becomes approximately equivalent to the overall variance, $\text{var}_{mpq}(\hat{\theta}_I - \theta)$, when the overall bias is negligible. The first, second and third terms on the right-hand side of (5.3) are referred to as the sampling variance, the nonresponse variance and the mixed component respectively. The sum of the last two terms can be called the nonresponse component since these terms would disappear if there were no nonresponse. The nonresponse component is simply the difference between the overall MSE/variance and the sampling variance. In what follows, we develop an estimator for each of these three terms.

5.1 Estimation of the sampling variance

Let $v(y)$ be a p -unbiased estimator of $\text{var}_p(\hat{\theta})$ that would be used under complete response. The typical Horvitz-Thompson estimator is

$$v(y) = \sum_{k \in s} \sum_{l \in s} \frac{\pi_{kl} - \pi_k \pi_l}{\pi_{kl}} (w_k d_k y_k)(w_l d_l y_l), \quad (5.4)$$

where π_{kl} is the joint selection probability of units k and l . In the presence of nonresponse, $\hat{V}_{\text{ORD}} = v(y_*)$ is the naïve sampling variance estimator that treats the imputed values as true values, where y_* is the imputed y -variable; i.e., $y_{*k} = y_k$, for $k \in s_r$, and $y_{*k} = y_k^*$, for $k \in s_m$.

Särndal (1992) proposed the following mpq -unbiased estimator of the sampling variance $V_{\text{SAM}} = E_m \text{var}_p(\hat{\theta})$:

$$\hat{V}_{\text{SAM}} = \hat{V}_{\text{ORD}} + \hat{V}_{\text{DIF}},$$

where \hat{V}_{DIF} is an m -unbiased estimator of $V_{\text{DIF}} = E_m(v(y) - \hat{V}_{\text{ORD}} | s, s_r)$. Unfortunately, the expression for \hat{V}_{DIF} is usually tedious to derive, and it is even more so when composite imputation is used.

Beaumont and Bocci (2009) simplified Särndal's derivations by conditioning on \mathbf{Y}_r , the vector containing the responding y -values. More explicitly, let $V_{\text{DIF}}^C = E_m(v(y) - \hat{V}_{\text{ORD}} | s, s_r, \mathbf{Y}_r)$ and \hat{V}_{DIF}^C be an m -unbiased estimator of V_{DIF}^C ; i.e., $E_m(\hat{V}_{\text{DIF}}^C | s, s_r, \mathbf{Y}_r) = V_{\text{DIF}}^C$. Our mpq -unbiased sampling variance estimator is $\hat{V}_{\text{SAM}}^C = \hat{V}_{\text{ORD}} + \hat{V}_{\text{DIF}}^C$. Since \hat{V}_{ORD} is a constant when conditioning on s , s_r and \mathbf{Y}_r , \hat{V}_{SAM}^C can simply be obtained by estimating $E_m(v(y) | s, s_r, \mathbf{Y}_r)$. If (5.4) is used,

$$E_m(v(y) | s, s_r, \mathbf{Y}_r) = v(y_*^\mu) + \sum_{k \in s_m} (1 - \pi_k) w_k^2 d_k \sigma_k^2, \quad (5.5)$$

where $y_{*k}^\mu = y_k$, for $k \in s_r$, and $y_{*k}^\mu = \mu_k$, for $k \in s_m$. An estimator \hat{V}_{SAM}^C of (5.5) is obtained by replacing the unknown mean μ_k and unknown variance σ_k^2 in (5.5) by m -unbiased (or at least m -consistent) estimators $\hat{\mu}_k$ and $\hat{\sigma}_k^2$. This estimator is easy to compute provided a software package that treats the complete response case is available to obtain the first term on the right-hand side of (5.5). The general formula (5.5) can be used for every imputation strategy. The only difference between different imputation strategies lies in the choice of the imputation model and the estimators $\hat{\mu}_k$ and $\hat{\sigma}_k^2$.

5.2 Estimation of the nonresponse variance

An mpq -unbiased estimator of the nonresponse variance $V_{\text{NR}} = E_{pq} E_m \{(\hat{\theta}_I - \hat{\theta})^2 | s, s_r\}$ is obtained by finding an m -unbiased estimator of

$$E_m\{(\hat{\theta}_I - \hat{\theta})^2 | s, s_r\} = \text{var}_m\{(\hat{\theta}_I - \hat{\theta}) | s, s_r\} + B_m^2. \quad (5.6)$$

Using $\hat{\theta}_I$ defined in the first equation of (3.3), the nonresponse error with composite imputation can be decomposed into J components:

$$\hat{\theta}_I - \hat{\theta} = \sum_{j=1}^J (\Omega_I^{(j)} - \Omega^{(j)}),$$

where $\Omega^{(j)} = \sum_{k \in s_m^{(j)}} w_k d_k y_k$. Each of these J components, $\Omega_I^{(j)} - \Omega^{(j)}$, is associated with a different imputation method. Since y_k^* only involves observed y -values, $\Omega_I^{(j)} = \sum_{k \in s_m^{(j)}} w_k d_k y_k^*$ only involves observed y -values as well and thus $\Omega_I^{(j)}$ and $\Omega^{(j)}$ are independent under the model. Therefore, the model variance of the nonresponse error can be written as

$$\begin{aligned} \text{var}_m\{(\hat{\theta}_I - \hat{\theta}) | s, s_r\} &= \sum_{i=1}^J \sum_{j=1}^J \text{cov}_m(\Omega_I^{(i)}, \Omega_I^{(j)} | s, s_r) \\ &\quad + \sum_{j=1}^J \text{var}_m(\Omega^{(j)} | s, s_r). \end{aligned} \quad (5.7)$$

Note that the covariances $\text{cov}_m(\Omega_I^{(i)}, \Omega_I^{(j)} | s, s_r)$, for $i \neq j$, are not necessarily negligible because some observed y -values can be used for more than one imputation method.

The derivations of the model variance (5.7) could be quite involved when several imputation methods are used because of the non-negligible covariances. The algebra can be greatly simplified for linear imputation methods. By using the second equation given in (3.3), the nonresponse error can be expressed as

$$\hat{\theta}_I - \hat{\theta} = W_{0d}^{(+)} + \sum_{k \in s_r} W_{dk}^{(+)} y_k - \sum_{k \in s_m} w_k d_k y_k. \quad (5.8)$$

Since the nonresponse error is linear in the y -values, its model variance is given by

$$\text{var}_m\{(\hat{\theta}_I - \hat{\theta}) | s, s_r\} = \sum_{k \in s_r} (W_{dk}^{(+)})^2 \sigma_k^2 + \sum_{k \in s_m} w_k^2 d_k \sigma_k^2. \quad (5.9)$$

If the model bias B_m is negligible, an mpq -unbiased estimator \hat{V}_{NR} of the nonresponse variance V_{NR} is obtained by replacing σ_k^2 in (5.9) by an m -unbiased (and m -consistent) estimator $\hat{\sigma}_k^2$. If the model bias is not negligible, it can be estimated by an m -consistent estimator \hat{B}_m and, using equation (5.6), the nonresponse variance estimator \hat{V}_{NR} can be replaced by $\hat{V}_{\text{NR}} + \hat{B}_m^2$. Note that \hat{B}_m^2 is m -consistent for B_m^2 provided that \hat{B}_m is m -consistent for B_m . The estimator \hat{B}_m can be found by using (5.8) and writing the model bias as

$$\begin{aligned} B_m &= E_m(\hat{\theta}_I - \hat{\theta} | s, s_r) \\ &= W_{0d}^{(+)} + \sum_{k \in s_r} W_{dk}^{(+)} \mu_k - \sum_{k \in s_m} w_k d_k \mu_k. \end{aligned} \quad (5.10)$$

The estimator \hat{B}_m is obtained by replacing μ_k in (5.10) by an m -consistent estimator $\hat{\mu}_k$.

5.3 Estimation of the mixed component

An mpq -unbiased estimator of the mixed component

$$V_{\text{MIX}} = 2E_{pq}E_m\{(\hat{\theta}_I - \hat{\theta})(\hat{\theta} - \theta) | s, s_r\}$$

is obtained by finding an m -unbiased estimator of

$$\begin{aligned} 2E_m\{(\hat{\theta}_I - \hat{\theta})(\hat{\theta} - \theta) | s, s_r\} &= \\ 2\text{cov}_m\{(\hat{\theta}_I - \hat{\theta}), (\hat{\theta} - \theta) | s, s_r\} &+ 2B_mE_m\{(\hat{\theta} - \theta) | s, s_r\}. \end{aligned} \quad (5.11)$$

Since both the nonresponse error and the sampling error are linear in the y -values, using (5.8) we obtain:

$$\begin{aligned} 2\text{cov}_m\{(\hat{\theta}_I - \hat{\theta})(\hat{\theta} - \theta) | s, s_r\} &= \\ 2\sum_{k \in s_r} W_{dk}^{(+)}(w_k - 1)d_k \sigma_k^2 - 2\sum_{k \in s_m} w_k(w_k - 1)d_k \sigma_k^2. \end{aligned} \quad (5.12)$$

If the model bias B_m is negligible, an mpq -unbiased estimator \hat{V}_{MIX} of the mixed component V_{MIX} is obtained by replacing σ_k^2 in (5.12) by an m -unbiased (and m -consistent) estimator $\hat{\sigma}_k^2$. Note that the mixed component is not necessarily negligible (Brick, Kalton and Kim 2004) and, moreover, it has been found to often be negative in practice.

If the model bias B_m is not negligible, it may not be possible to easily estimate the second component on the right-hand side of (5.11). The reason is that $E_m\{(\hat{\theta} - \theta) | s, s_r\}$ involves knowing $\mathbf{x}_k^{\text{obs}}$ as well as the domain indicator variable d for the nonsampled portion of the population; this information may not be available. This problem can be bypassed by changing the inferential framework. The full multivariate distribution between y , \mathbf{x} and d can be modeled instead of conditioning on d and \mathbf{x}^{obs} . We did not implement this idea in SEVANI because it leads to a more complex modeling task and makes it difficult to obtain a general variance expression that is easy to implement. Ignoring the second component on the right-hand side of (5.11) should not be of great concern in practice when the model bias is not too large. In section 5.4, we provide a diagnostic that can be helpful for determining whether the model bias is important or not.

The mixed component can also be written as

$$\begin{aligned} V_{\text{MIX}} &= 2E_{pq} E_m \{(\hat{\theta}_I - \hat{\theta})(\hat{\theta} - \theta) | s, s_r\} \\ &= 2E_{pq} [\text{cov}_m \{(\hat{\theta}_I - \hat{\theta}), (\hat{\theta} - \theta) | s, s_r\}] \\ &\quad + 2E_p [E_q(B_m | s) E_m \{(\hat{\theta} - \theta) | s\}]. \end{aligned}$$

Expression (5.12) can therefore be used to obtain an estimator of V_{MIX} provided that $E_q(B_m | s)$ is negligible. This is a weaker assumption than requiring B_m to be negligible since this assumption is satisfied when either B_m or $E_q(\hat{\theta}_I - \hat{\theta} | s)$ is negligible. For instance, in our earlier example, B_m may not be negligible but, if $d_k = 1$ and $\omega_k^{(1)} = \omega_k^{(2)} = w_k$, $E_q(\hat{\theta}_I - \hat{\theta} | s) \approx 0$ under uniform non-response (see Sitter and Rao 1997).

5.4 Estimation of the overall MSE/variance

The overall MSE, or overall variance if the overall bias is negligible,

$$V_{\text{TOT}} = E_{mpq} (\hat{\theta}_I - \theta)^2 = V_{\text{SAM}} + V_{\text{NR}} + V_{\text{MIX}}$$

can be estimated by $\hat{V}_{\text{TOT}} = \hat{V}_{\text{SAM}}^C + \hat{V}_{\text{NR}} + \hat{V}_{\text{MIX}}$ if the model bias, B_m , is negligible. The nonresponse component estimator is $\hat{V}_{\text{NR}} + \hat{V}_{\text{MIX}}$. From a user's perspective, the estimator \hat{V}_{TOT} is of greater interest than its individual components. A user may nevertheless be interested in the estimator of the sampling variance, \hat{V}_{SAM}^C , or the ratio $\hat{V}_{\text{SAM}}^C / \hat{V}_{\text{TOT}}$. The latter estimates the contribution of the sampling variance to the overall variance.

As pointed out in section 5.2, if the model bias is not negligible, the nonresponse variance can be estimated by $\hat{V}_{\text{NR}} + \hat{B}_m^2$ instead of \hat{V}_{NR} . This leads to the overall MSE estimator $\hat{V}_{\text{TOT,ADJ}} = \hat{V}_{\text{SAM}}^C + (\hat{V}_{\text{NR}} + \hat{B}_m^2) + \hat{V}_{\text{MIX}}$.

A statistic that can be useful as a diagnostic to determine the magnitude of the model bias is either $|\hat{B}_m| / \sqrt{\hat{V}_{\text{TOT}}}$ or $|\hat{B}_m| / \sqrt{\hat{V}_{\text{TOT,ADJ}}}$. A large value of any of these two statistics may be an indication that the model bias is not negligible and that the composite imputation procedure should be questioned. The advantage of $|\hat{B}_m| / \sqrt{\hat{V}_{\text{TOT,ADJ}}}$ over $|\hat{B}_m| / \sqrt{\hat{V}_{\text{TOT}}}$ is that it is bounded; i.e.,

$$0 \leq |\hat{B}_m| / \sqrt{\hat{V}_{\text{TOT,ADJ}}} \leq 1.$$

5.5 Random regression imputation

A random regression residual e_k is sometimes added to the regression imputed value y_k^* to preserve the natural variability of the y -variable. We suggest that the random residuals e_k be generated independently with $E_*(e_k | s, s_r) = 0$ and $\text{var}_*(e_k | s, s_r) = \hat{\sigma}_k^2$, where the subscript $*$ indicates that the expectation and variance are taken with respect to the random imputation mechanism. This leads to

the imputed value $y_k^{*R} = y_k^* + r_k e_k$, with $r_k = 1$ if unit k has been imputed with a random residual added and $r_k = 0$ otherwise. The imputed estimator (2.1) with y_k^* replaced by y_k^{*R} is denoted by $\hat{\theta}_I^* = \hat{\theta}_I + \sum_{k \in s_m} w_k d_k r_k e_k$. Since $E_*(e_k | s, s_r) = 0$, adding a random residual does not introduce any bias in the imputed estimator. The overall MSE of $\hat{\theta}_I^*$ can be expressed as

$$E_{mpq^*} (\hat{\theta}_I^* - \theta)^2 = E_{mpq} (\hat{\theta}_I - \theta)^2 + E_{mpq} \text{var}_*(\hat{\theta}_I^* | s, s_r). \quad (5.13)$$

The first term on the right-hand side of (5.13) is estimated as in section 5.4. The second term is estimated by

$$\text{var}_*(\hat{\theta}_I^* | s, s_r) = \sum_{k \in s_m} w_k^2 d_k r_k \hat{\sigma}_k^2. \quad (5.14)$$

6. Simulation study

We conducted a Monte-Carlo simulation study to assess the methodology described in section 5. A bivariate population of $N = 400$ units was generated that contains an auxiliary variable x and a variable of interest y . For each population unit, the auxiliary variable was generated according to a gamma distribution with mean 48 and variance 768. The variable of interest y was generated conditionally on x from a gamma distribution with mean $1.5x$ and variance $16x$. Half of the population was randomly assigned a missing value to x . As no domain of interest was generated, θ is the overall population total of variable y .

Ten thousand samples were selected from this population using simple random sampling without replacement. We considered two sample sizes: $n = 100$ and $n = 250$. For each sample, nonresponse to variable y was generated independently from one unit to another with a nonresponse probability of 0.3. We used the same imputation strategy as in the example in section 2 with $\omega_l^{(1)} = 1$, for $l \in s_r^{(1)}$, and $\omega_l^{(2)} = 1$, for $l \in s_r$. Nonrespondents to variable y with an observed x -value were imputed by ratio imputation while those with a missing x -value were imputed by mean imputation.

The population y -values were kept fixed throughout the replications of the simulation experiment; each replication consisted of selecting a sample and then generating nonresponse to variable y . If we had strictly followed the theoretical development in section 5, we would have generated new y -values at each replication according to the imputation model. However, it is more common in the literature to fix the population y -values when conducting a simulation experiment. For instance, our simulation set-up is essentially the same as the one discussed in Rancourt, Lee and Särndal (1993), who also considered composite imputation.

We computed the Monte-Carlo sampling variance and overall MSE as $V_{\text{SAM}}^{\text{MC}} = \sum_{r=1}^R (\hat{\theta}_r - \theta)^2 / R$ and $V_{\text{TOT}}^{\text{MC}} = \sum_{r=1}^R (\hat{\theta}_{I,r} - \theta)^2 / R$ respectively, where the subscript r indicates that estimates are computed using the r^{th} replicate and $R = 10,000$. The Monte-Carlo relative bias of any estimator of V_{SAM} , say v_{SAM} , is computed as $\text{RB}(V_{\text{SAM}}) = \sum_{r=1}^R (v_{\text{SAM},r} - V_{\text{SAM}}^{\text{MC}}) / (V_{\text{SAM}}^{\text{MC}} R)$. Similarly, we computed the Monte-Carlo relative bias of an estimator of V_{TOT} , denoted as $\text{RB}(V_{\text{TOT}})$, and the Monte-Carlo relative bias of an estimator of $V_{\text{SAM}} / V_{\text{TOT}}$, denoted as $\text{RB}(V_{\text{SAM}} / V_{\text{TOT}})$. Finally, we computed the Monte-Carlo coverage rates of confidence intervals for θ with a 95% confidence level assuming that $\hat{\theta}_I$ is normally distributed.

The results of our simulation study are given in table 2. In the columns labeled SEVANI, the sampling variance, V_{SAM} , and the overall MSE, V_{TOT} , are estimated for each sample by $\hat{V}_{\text{SAM}}^{\text{C}}$ and $\hat{V}_{\text{TOT,ADJ}}^{\text{C}}$ respectively (see section 5.4). We have also obtained results by replacing $\hat{V}_{\text{TOT,ADJ}}^{\text{C}}$ by $\hat{V}_{\text{TOT}}^{\text{C}}$. We do not report these additional results in table 2 as they were quite close to those obtained with $\hat{V}_{\text{TOT,ADJ}}^{\text{C}}$. This suggests that the model bias B_m is not important in this case. In the columns labeled Naïve, both the sampling variance and the overall MSE are estimated by $\hat{V}_{\text{ORD}}^{\text{C}}$ (see section 5.1).

Table 2
Results of the simulation study

	$n = 100$		$n = 250$	
	SEVANI	Naïve	SEVANI	Naïve
$\text{RB}(V_{\text{SAM}})$	2.82%	-17.59%	3.02%	-17.68%
$\text{RB}(V_{\text{SAM}}/V_{\text{TOT}})$	8.30%	-	5.84%	-
$\text{RB}(V_{\text{TOT}})$	-5.07%	-40.68%	-2.66%	-52.89%
Coverage Rate	93.38%	86.20%	94.42%	81.80%

These results show that the methodology described in section 5 and implemented in SEVANI is better than the naïve variance estimator for the estimation of the components of variance and the construction of confidence intervals. The use of SEVANI leads to small Monte-Carlo relative biases and coverage rates close to the targeted nominal rate (95%). Our methodology is also useful for users who would like to estimate the contribution of the sampling variance to the overall MSE; *i.e.*, $V_{\text{SAM}} / V_{\text{TOT}}$. Note that $V_{\text{SAM}}^{\text{MC}} / V_{\text{TOT}}^{\text{MC}}$ is 71.98% for $n = 100$ and 57.23% for $n = 250$. Since $V_{\text{SAM}}^{\text{MC}} / V_{\text{TOT}}^{\text{MC}}$ is not close to 100% even for $n = 100$, the effects of nonresponse and imputation cannot be systematically ignored when estimating the overall MSE.

7. The reverse approach

Shao and Steel (1999) proposed a reverse approach to variance estimation developed to deal with composite imputation. They assumed that the overall bias is negligible and suggested the following decomposition of the overall variance:

$$E_{mpq}(\hat{\theta}_I - \theta)^2 = E_{mq} \text{var}_p(\hat{\theta}_I | U_r) + E_{mq} \{E_p(\hat{\theta}_I | U_r) - \theta\}^2, \quad (7.1)$$

where U_r is a conceptual population of respondents. The inner expectation and variance in the right side of (7.1) are taken with respect to the sampling design. Unfortunately, the imputed estimator $\hat{\theta}_I$ is generally not linear with respect to the sampling design even though it is linear with respect to the observed y -values. Therefore, the imputed estimator $\hat{\theta}_I$ is typically linearized (*e.g.*, Shao and Steel 1999; Kim and Rao 2009). More explicitly, the quantities $\phi_{0k}^{(j)}$ and $\phi_{lk}^{(j)}$ often depend on the sample in a nonlinear way; *e.g.*, this is true with linear regression imputation (see the example at the end of section 3) and donor imputation. It is not always straightforward to account for the sampling variability of $\phi_{0k}^{(j)}$ and $\phi_{lk}^{(j)}$ when using (7.1). For example, there is no literature on the use of the reverse approach to estimate the variance under nearest-neighbour imputation. Moreover, since each composite imputation strategy yields its own linearized imputed estimator, it is not an easy task to implement this methodology in a generalized software package.

Using our approach, the inner expectation in the expressions for the nonresponse variance,

$$V_{\text{NR}} = E_{pq} E_m \{(\hat{\theta}_I - \hat{\theta})^2 | s, s_r\},$$

and the mixed component,

$$V_{\text{MIX}} = 2E_{pq} E_m \{(\hat{\theta}_I - \hat{\theta})(\hat{\theta} - \theta) | s, s_r\},$$

are taken with respect to the imputation model (conditionally on s and s_r). The imputed estimator is linear and the derivations are straightforward because the quantities $\phi_{0k}^{(j)}$ and $\phi_{lk}^{(j)}$ are constructed without using the y -values. The estimation of the sampling variance, $V_{\text{SAM}} = E_m \text{var}_p(\hat{\theta})$, does not involve these two quantities (see equation 5.5); thus, their possible non-linearity with respect to the sampling design does not cause any difficulty. This implies that nearest-neighbour imputation can be easily handled with our approach (see Beaumont and Bocci 2009).

It is for all the above reasons that we believe that the reverse approach might be more cumbersome to implement in a generalized software package than our approach. This

does not mean that the reverse approach is not useful. Indeed, both approaches lead to identical variance estimators when a census is conducted. Beaumont, Haziza and Bocci (2011) showed that they also lead to identical variance estimators under auxiliary value imputation (because $\varphi_{0k}^{(j)}$ and $\varphi_{ik}^{(j)}$ do not depend on s and s_r). Both approaches depend on the correct specification of the imputation model and no approach is expected to systematically outperform the other.

The reverse approach may have a practical advantage over our approach when the sampling fraction is negligible. In such case, Shao and Steel (1999) showed that the second component on the right side of (7.1) can be neglected. The first component is estimated by finding a design-based estimator of $\text{var}_p(\hat{\theta}_I | U_r)$. If a replication variance estimation technique (e.g., the jackknife or the bootstrap) is chosen for the estimation of $\text{var}_p(\hat{\theta}_I | U_r)$, the whole approach becomes quite attractive and practical. Also, it does not depend on the validity of the imputation model; in particular, the correct specification of the model variance σ_k^2 . The jackknife variance estimators of Rancourt, Lee and Särndal (1993) and Sitter and Rao (1997) can be justified by this approach.

8. Conclusion

Our methodology for composite imputation has been implemented in version 2 of SEVANI because of its ease of implementation and generality. It works for most imputation methods used in practice, as most imputation methods are linear. The variance computations are the same for every composite imputation strategy once the quantities $W_{0d}^{(+)}$, $W_{dk}^{(+)}$, $\hat{\mu}_k$ and $\hat{\sigma}_k^2$ have been computed. This eases the development of a generalized system.

Although we have focused on the estimation of a domain total using the Horvitz-Thompson estimator, SEVANI can also deal with domain means and calibration estimators. Parametric and nonparametric methods of estimating μ_k and σ_k^2 are also available. Greater detail can be found in the Methodology Guide of SEVANI (Beaumont, Bissonnette and Bocci 2010) available upon request from the authors.

Acknowledgements

We would like to thank the reviewers for their comments. We would also like to thank Mike Hidioglou, Eric Rancourt and Cynthia Bocci from Statistics Canada for their suggestions and discussions on the topic. All these comments contributed to improve the paper.

References

- Beaumont, J.-F., Bissonnette, J. and Bocci, C. (2010). SEVANI, version 2.3, Methodology Guide. Internal report, Methodology Branch, Statistics Canada.
- Beaumont, J.-F., and Bocci, C. (2009). Variance estimation when donor imputation is used to fill in missing values. *Canadian Journal of Statistics*, 37, 400-416.
- Beaumont, J.-F., Haziza, D. and Bocci, C. (2011). On variance estimation under auxiliary value imputation in sample surveys. *Statistica Sinica*, 21, 515-537.
- Brick, J.M., Kalton, G. and Kim, J.K. (2004). Variance estimation with hot deck imputation using a model. *Survey Methodology*, 30, 57-66.
- Deville, J.-C., and Särndal, C.-E. (1994). Variance estimation for the regression imputed Horvitz-Thompson estimator. *Journal of Official Statistics*, 10, 381-394.
- Felix, P., and Rancourt, E. (2001). Applications of Variance due to Imputation in the Survey of Employment, Payrolls and Hours. Methodology Branch Working Paper, Statistics Canada, BSMD-2001-009E.
- Haziza, D. (2009). Imputation and inference in the presence of missing data. In *Handbook of Statistics, Sample Surveys: Theory, Methods and Inference*, (Eds., D. Pfeffermann and C.R. Rao). Amsterdam: Elsevier BV, 29A, 215-246.
- Hidioglou, M.A. (1989). Unpublished handwritten notes kindly shared with us by the author.
- Kim, J.-K., and Rao, J.N.K. (2009). Unified approach to linearization variance estimation from survey data after imputation for item nonresponse. *Biometrika*, 96, 917-932.
- Lee, H., Rancourt, E. and Särndal, C.-E. (2001). Variance estimation from survey data under single imputation. In *Survey Nonresponse*, (Eds., R.M. Groves, D.A. Dillman, J.L. Eltinge and R.J.A. Little). New-York: John Wiley & Sons, Inc., 315-328.
- Rancourt, E., Lee, H. and Särndal, C.-E. (1993). Variance estimation under more than one imputation method. In *Proceedings of the International Conference on Establishments Surveys*, June 1993, Buffalo, American Statistical Association, 374-379.
- Rubin, D.B. (1987). *Multiple Imputation for Nonresponse in Surveys*. New-York: John Wiley & Sons, Inc.
- Särndal, C.-E. (1992). Methods for estimating the precision of survey estimates when imputation has been used. *Survey Methodology*, 18, 241-252.
- Shao, J., and Steel, P. (1999). Variance estimation for survey data with composite imputation and nonnegligible sampling fractions. *Journal of the American Statistical Association*, 94, 254-265.
- Sitter, R.R., and Rao, J.N.K. (1997). Imputation for missing values and corresponding variance estimation. *Canadian Journal of Statistics*, 25, 61-73.

Alternative demographic sample designs being explored at the U.S. Census Bureau

Patrick E. Flanagan and Ruth Ann Killion¹

1. Introduction

The United States (U.S.) Census Bureau Demographic Survey Sample Redesign Program, among other things, is responsible for research into improving the designs of U.S. demographic surveys, particularly focused on the design of survey sampling. Historically, the research into improving sample design has been restricted to the “mainstream” methods like basic stratification, multi-stage designs, systematic sampling, probability-proportional-to-size sampling, clustering, and simple random sampling. Over the past thirty years or more, we have increasingly faced reduced response rates and higher costs coupled with an increasing demand for more data on all types of populations. More recently, dramatic increases in computing power and availability of auxiliary data from administrative records have indicated that we may have more options than we did when we established our current methodology. Thus, we began an initiative to explore alternative sampling methods.

2. History of innovation in demographic survey sampling at the U.S. Census Bureau

The U.S. Census Bureau was created by the Permanent Census Act of 1902. Up until the late 1930s, the U.S. Census Bureau’s demographic work was mostly focused on the logistics of running each decennial census and a myriad of special censuses. After the 1930 decennial census, the Census Bureau began research into sampling using the census data (Stephan 1948).

Then, in 1937, the Census Bureau took its first major step into sample survey sampling with the 1937 Enumerative Check Census of Unemployment, which used a cluster sample of counties in support of a register census of the unemployed (Dedrick 1938). About the same time, the Census Bureau brought in sampling experts (*e.g.*, W. Edwards Deming and Frederick Stephan) in its decennial census expansion to assist in designing a sample survey in conjunction with the 1940 Decennial Census using a five percent systematic sample (Stephan, Deming and Hansen 1940). In 1942, the Sample Survey of Unemployment was moved from the Works Progress Administration to the Census Bureau. This survey was already a three-stage sample with county primary sampling units (PSUs), systematic sampling of blocks, and sampling listed housing units in

stage three (Frankel and Stock 1942). After its transfer to the Census Bureau (and a name change to the Monthly Report on the Labor Force (MRLF)), it was extensively redesigned in 1943, dramatically improving its efficiency using larger primary sampling units (PSUs) and probability proportionate to size for selection (Duncan and Shelton 1978). Later the survey was changed to improve month-to-month and year-to-year comparisons using a more complex overlapping sample approach in which a given household remains in sample for four months, is out of the survey for eight months and then is back into the sample for four months. Its name was also changed in 1947 to the Current Population Survey (CPS). Still, the basic sampling concept remained multi-stage sample design with county or county group PSUs. It remains that way to present though there are vast differences in the within-PSU sampling methods (U.S. Bureau of Labor Statistics and U.S. Census Bureau 2006). Over the last 60 years, the U.S. Census Bureau has designed many additional demographic surveys. Some of those surveys use the same two-stage design idea used in the CPS, like the Consumer Expenditures Surveys, the Survey of Income and Program Participation, the National Crime Victimization Survey, and the National Health Interview Survey. Some others are two-stage with selection of a list source followed by sampling from the lists like the Schools and Staffing Survey, the Private School Survey, and the Survey of Inmates of Local Jails. Still other are stratified samples from a sampled frame, such as the National Survey of College Graduates that has sampled from the Decennial Census Long Form, and the American Time Use Survey that samples from the CPS. In the early 1990s, The U.S. Census Bureau initiated the development of the use of continuous measurement as a possible replacement for the Decennial Census Long Form. Those efforts have since evolved into the current American Community Survey, which, starting 2010, will provide continual mid-decade estimates down to the block group level. The Census Bureau’s goal for improving our sampling methodology to the present leads us to explore alternative sample designs.

3. Alternative survey sample design seminar series

The exploration into alternative methods of sampling began with an initial seminar series that was held at the U.S.

1. Patrick E. Flanagan and Ruth Ann Killion, U.S. Census Bureau. E-mail: Patrick.e.flanagan@census.gov.

Census Bureau. It consisted of three seminar presentations of such methods covering the statistical bases of the methods and their limitations, especially when applied to the types of demographic surveys conducted by the U.S. Census Bureau. Each presentation also included discussant comments by Professor Jean Opsomer from Colorado State University. Three articles were then developed providing greater detail on each topic and a final discussant article covering the three subjects.

- On 26 September 2007, Professor Steven K. Thompson from Simon Fraser University gave a presentation on his research into network sampling, spatial sampling, and adaptive sampling.
- On 9 January 2008, Professor Sharon Lohr from Arizona State University gave a presentation on her research into sampling using overlapping frames.
- On 4 June 2008, Professor Yves Tillé from University of Neuchatel gave a presentation on his research into balanced sampling.

The articles resulting from this project that follow are:

“Adaptive network and spatial sampling,” by Steven Thompson;

“Alternative survey sample designs: Sampling with multiple overlapping frames,” by Sharon Lohr;

“Ten years of balanced sampling with the cube method: An appraisal,” by Yves Tillé; and

“Innovations in survey sampling design: Discussion of three contributions presented at the U.S. Census Bureau,” by Jean Opsomer.

4. Next steps

Following these three presentations, it was decided to conduct further research into these methods and their application to either existing U.S. Census Bureau Demographic surveys or to potential new surveys. There is already an urgent need for using multiple overlapping frames methods applied to the National Survey of College Graduates to deal with an old-cohort/new-cohort problem and a possible use of state hunting and fishing license registries as a second frame for the Fishing, Hunting, and Wildlife-Associated Recreation survey. We have plans to look at balanced sampling, particularly for selecting

geographic primary sampling units. Lastly, the methods of adaptive sampling have the potential for us to accept surveys that we traditionally have not taken on, as well as providing a lower cost alternative for surveys that meet certain criteria.

5. Summary

This exploration into these three areas of alternative sample designs is just the beginning of our seminar series and of our intentions to explore methods to improve our demographic survey sample design methods. Future anticipated subjects include alternative listing methods, Kish's half-open interval approach to growth updates and coverage improvement, responsive survey designs, rejective sampling procedures, and model-assisted sampling.

Acknowledgements

This report is released to inform interested parties of research and to encourage discussion. The views expressed on statistical, methodological, technical, or operational issues are those of the authors and not necessarily those of the U.S. Census Bureau.

References

- Dedrick, C.L. (1938). *Census of unemployment 1937: Principle findings of the enumerative check census*. U.S. Bureau of the Census.
- Duncan, J.W., and Shelton, W.C. (1978). *Revolution in United States Government Statistics 1926 – 1976*. U.S. Department of Commerce.
- Frankel, L.R., and Stock, J.S. (1942). On the sample survey of unemployment. *Journal of the American Statistical Association*, 37, 77-80.
- Stephan, F.F., Deming, W.E. and Hansen, M.H. (1940). The sampling procedure of the 1940 population census. *Journal of the American Statistical Association*, 35, 615-630.
- Stephan, F.F. (1948). History of the uses of modern sampling procedures. *Journal of the American Statistical Association*, 43, 12-39.
- U.S. Bureau of Labor Statistics and U.S. Census Bureau (2006). *Design and Methodology: Current Population Survey*.

Adaptive network and spatial sampling

Steve Thompson¹

Abstract

This paper describes recent developments in adaptive sampling strategies and introduces new variations on those strategies. Recent developments described included targeted random walk designs and adaptive web sampling. These designs are particularly suited for sampling in networks; for example, for finding a sample of people from a hidden human population by following social links from sample individuals to find additional members of the hidden population to add to the sample. Each of these designs can also be translated into spatial settings to produce flexible new spatial adaptive strategies for sampling unevenly distributed populations. Variations on these sampling strategies include versions in which the network or spatial links have unequal weights and are followed with unequal probabilities.

Key Words: Network sampling; Snowball sampling; Random walk; Markov chain; Adaptive web sampling.

1. Introduction

An adaptive sampling design is a procedure for selecting the sample in which the probabilities of selecting the set of sample units from the population depend on values of the variable of interest observed during the survey. In a spatial setting, adaptive sampling is exemplified by a survey in which, whenever a unit in the sample is observed to have an unusually high or otherwise interesting value of the variable of interest, nearby units may be added to the sample. In a network setting such as a socially networked human subpopulation, a link-tracing design may be used to adaptively follow social links from sample individuals to locate and add additional members of the subpopulation to the sample.

In spatial settings the development of adaptive designs has been motivated by such problems as estimating the abundance of rare, clustered plant and animal species, assessment of unevenly distributed environmental pollutants, and surveys of geographically clustered subpopulations of people. In network settings the development of adaptive network sampling designs has been motivated by problems in sampling people with rare diseases, sampling hidden populations such as those at high risk for HIV/AIDS or other epidemics, and sampling through computer and communications networks.

Zacks (1969) and Basu (1969) recognized that in most cases the optimal sampling would in principle be an adaptive one. With a Bayes model for the population, at any step part way through a sampling procedure, one can do as well or better than a conventional design by selecting the remaining sample to give the lowest mean square error conditional on the observed sample values so far. The overall mean square error is the expected value of the conditional mean square error. The underlying mathematical principle is that the integral of the minimum of a set of functions is smaller, or not larger than, the minimum of the

integrals. Results on optimal adaptive strategies are described and extended in Thompson and Seber (1996) and exemplified in Chao and Thompson (2001).

In spite of the early theoretical results and motivation from field surveys, the importance of adaptive designs was not widely recognized for several decades in either theory or practice. The practical importance of adaptive sampling strategies became evident as statistical thinking was brought to bear on problems in natural resource management and environmental protection. The development of adaptive link-tracing designs for reaching hidden human populations has attained strategic importance for such problems as understanding and alleviating the global HIV epidemic. In addition, new interest in adaptive sampling methods is being spurred by problems of expense and effort in social surveys of all types.

Adaptive designs such as those described in this paper often serve as high yield designs in that sample values of variables of interest tend to be higher on average than population means of the same variables. Although this is often a desired characteristic itself in studies of rare populations, simple sample data summaries such as sample means and sample proportions are generally not good estimates of population means or proportions. Instead, effective design-based and model-based estimators of population quantities have been developed for use with adaptive designs.

With design-based estimators, properties such as unbiasedness or consistency depend solely on the way the sample is selected and not on assumptions about what the population may be like. Model-based estimators such as maximum likelihood or Bayes estimators on the other hand require use of a statistical model, usually involving unknown parameter values, describing the population of interest. Design-based estimators for adaptive designs are described in Thompson and Seber (1996), Thompson (2006a, b), and earlier papers.

1. Steve Thompson, Simon Fraser University. E-mail: Thompson@stat.sfu.ca.

Basic results for model-based approaches to inference with adaptive designs were given in Thompson and Seber (1996), which showed that likelihood-based methods such as maximum likelihood and Bayes inference would be more effective than other model based approaches (for example, the linear unbiased prediction approach) with adaptive designs. Maximum likelihood estimation and the likelihood based approach more generally with link-tracing designs were described in Thompson and Frank (2000). Bayes estimation with link tracing designs was used in Chow and Thompson (2003). A method combining model and design based features was used in Felix-Medina and Thompson (2004). Bayes estimation using Markov chain Monte Carlo (MCMC) with adaptive web sampling designs is described in Kwanisai (2005, 2006).

2. Adaptive sampling in network settings

A population has network structure if there are links or relationships between any of the units in the population. Mathematically, such a population is described as a graph, consisting of a set of nodes and a set of edges or arcs between nodes. More generally, each relationship between a pair of nodes may have a weight denoting the strength of value associated with the relationship.

Human populations have an inherent network structure arising from social relationships. As will be noted later, spatial relationships also give a network structure to many populations. Network populations also arise in computing networks, communications, gene regulation and metabolic networks.

Network structure in populations is important for two reasons. First, the network relationships may be of interest in themselves to researchers. For example, with contagious disease epidemics it is important to know the nature and pattern of the social contacts through which the disease spreads. Second, the network structure can be used to help in obtaining a sample from a population that is otherwise difficult to sample. For example, in the study of hidden populations at risk for HIV/AIDS, including drug injectors, commercial sex workers, and others, often the only way in many cases to obtain a sample large enough for the study is to follow social links from initial sample individuals to find more members of the hidden population.

Most network sampling designs which follow links are inherently adaptive in that the link values used in the selection are variables of interest that are generally not known prior to the survey. Further, in some studies it may be of interest to follow links with higher probability from sample individuals with high values of variables associated with behavioral risk.

A class of designs called *multiplicity sampling* or simply *network sampling* was introduced by Birnbaum and Sirken (1965), along with design-unbiased estimators of population quantities. The approach was developed further by Sirken (1970, 1972a, b) and others and is described in Thompson (2002). In these designs the units on which observations are made are obtained by first selecting “selection units”, to which the observational units are linked. Motivation for these strategies came from problems in public health in which commonly used estimates were found to be biased because of the unequal numbers of such links. The simplest of the unbiased estimators in terms of computations was the “multiplicity estimator” which simply divided the observed value of a variable of interest measured on an observational unit by its “multiplicity”, the number of selection units to which it is linked. Horvitz-Thompson estimators for the strategy were also introduced. The following decades saw many variations on this strategy published in the statistics and substantive literatures.

In *snowball sampling* an initial sample of nodes is selected by some design such as simple random sampling, and every link out is followed to add connected nodes to the sample. This process is continued for a specified number of steps, or “waves”. More generally, a subsample such as a fixed number of links are followed at each wave. Frank (1971, 1977a, b, 1978a, b, 1979) framed the problem as one of sampling in graphs worked out design-based estimators for many cases of snowball designs including designs with unequal initial selection probabilities and estimators for population quantities such as totals and means of variables associated with nodes or individuals, as well as of population link quantities such as mean degree, where degree of a node is defined as the number of links out (or in) from that node. Frank and Snijders (1994) introduced a number of design-based and model-based estimators for one wave snowball designs motivated by the problem of estimating the number of injection drug users in a city.

In a *random walk* design an initial node is selected at random. From the links out from that node one link is selected at random and followed to add the connected node to the sample. This process is continued for a specified number of waves, with one unit selected at each wave. If the sampling is with replacement the design is a Markov chain, with the state of the chain at each step being the identity of the node selected at that step. Properties of such designs, cast as Markov chains in graphs, such as the limiting or stationary probabilities were examined in the statistics and probability literatures (Lovász 1993). Random walk designs were introduced into the social network literature by Klov Dahl (1989) with the motivation of reaching into a hidden human population farther away from the initial sample than possible with the same sample size using a

snowball design. In the computing science literature, Brin and Page (1998) used the concept of a stationary distribution of a random walk in a graph in developing a search engine and web page ranking algorithm, evoking the metaphor of a “random surfer” to describe the process of a random walk following hyper-links from web page to web page.

Heckathorn (1997, 2002) and Salganik and Heckathorn (2004) described a sampling methodology referred to as “respondent driven sampling” in which members of a hidden population were motivated to recruit other members of the population into the sample using a system of coupons. A simple estimator of population totals and means, in which each observation is weighted by the reciprocal of that person’s degree, was used with these designs based on the limiting distribution of a with-replacement random walk in a network having symmetric links and a single connected component. The coupon-based methodologies developed with these designs have proven to be highly effective in recruiting samples of substantial size from hidden populations in a number of settings.

The notational setup for sampling in networks follows. There is a population of units or nodes labeled $1, 2, \dots, N$ with associated variables of interest y_1, y_2, \dots, y_N . Associated with each pair of nodes (i, j) is a link-indicator or weight, so that the collection $\{w_{ij}; i, j = 1, \dots, N\}$ are variables of interest associated with pairs of nodes.

In the network context a sample s is a subset $s^{(1)}$ of nodes and a subset $s^{(2)}$ of the pairs of nodes, that is, $s = (s^{(1)}, s^{(2)})$. Thus the sample consists of a sample of nodes, on which node variables y of interest are recorded, and a sample of pairs of nodes, for which the values of relationship variables w are recorded.

Figure 1 shows a network-structured population which will be used to illustrate some of the network sampling designs described in this paper. In terms of a human population with social network structure, the red or dark colored nodes could represent individuals with high values of variables of interest, for example indicating a risk-related behavior such as injecting drug use. The light colored or yellow nodes would represent the individuals without the high-risk characteristic of interest. The links between individuals would represent social relationships such as having meals together, drug-using relationships, or sexual contacts.

Figure 2 shows an initial simple random sample of five nodes selected from the network-structured population. A one-wave snowball sample selected by following every link out from the initial sample is shown in Figure 3, and a two-wave snowball sample from the same initial sample is shown in Figure 4. Note that with a fixed number of waves, a snowball sample can grow very fast.

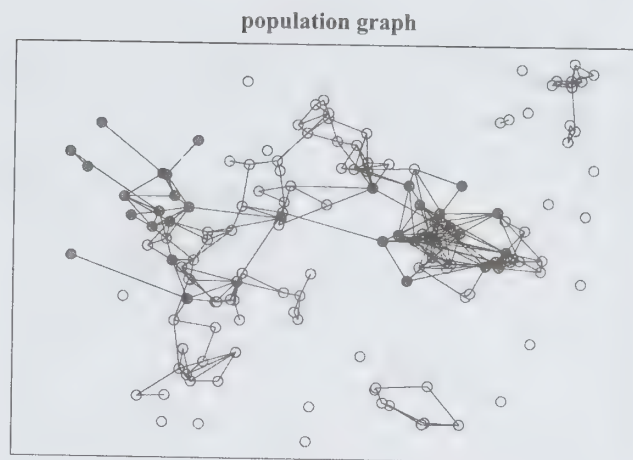


Figure 1 A population with network structure

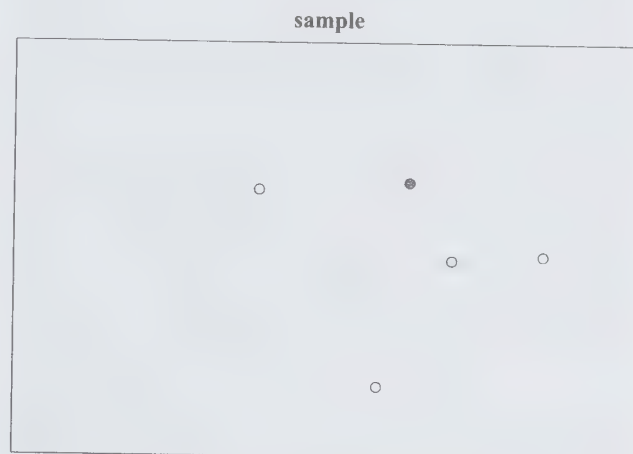


Figure 2 A random sample of nodes

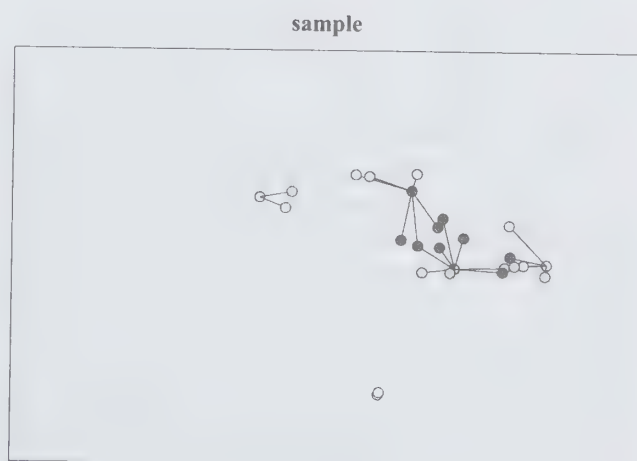


Figure 3 One-wave snowball sample

sample

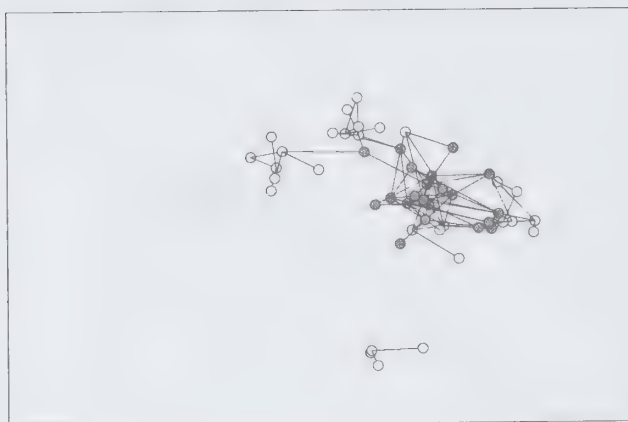


Figure 4 Two-wave snowball sample

With a snowball sampling designs and many other link-tracing designs, sample data summaries such as a sample mean or sample proportions are not good estimators of the analogous population characteristics. The reason is that under the design different units have different probabilities of selection, dependent on the population link structure. Figure 5 shows the population with the size of each node proportional to the probability of selecting that node. Since high-risk individuals tend to have more links hence higher probabilities of inclusion in the sample, the sample mean would tend to overestimate the population mean. In the same way, the average degree of such a sample would tend to overestimate the mean degree of the population network.

With the one-wave snowball design in a setting with symmetric links the inclusion probabilities for sample nodes can be easily calculated as proportional to the node degrees. With asymmetric links or with snowball designs of more than one wave it is not in general possible to calculate node inclusion probabilities from the sample data. Methods for calculating design-unbiased estimators of population node and link characteristics with such designs are described in the section on adaptive web sampling later in this paper.

Figure 6 shows a snowball sample from this same network population starting with one randomly selected unit. Since the population consists of more than a single connected component a strict random walk design would be stuck in whatever component it started in. It is therefore desirable to provide in the design some small probability at each step of selecting the next unit by simple random sampling or some other conventional design, or at least allowing a random jump whenever a walk is found to be stuck in a component.

Figure 7 shows the stationary selection probabilities for the random walk through the network shown. Although these probabilities in this population are not simply proportional to node degrees it can be seen that nodes with

high degree do tend to have high selection probabilities. Also, since high risk individuals in this population tend to have high selection probability under this design, sample summaries such as sample mean and sample proportion are not unbiased estimators of population means and proportions. For unbiased estimates the methods of later sections of this paper would have to be used.

one-wave selection probabilities

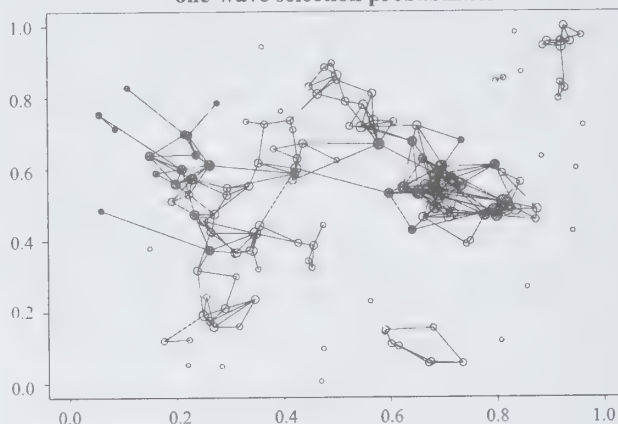


Figure 5 One-wave snowball sample selection probabilities

walk

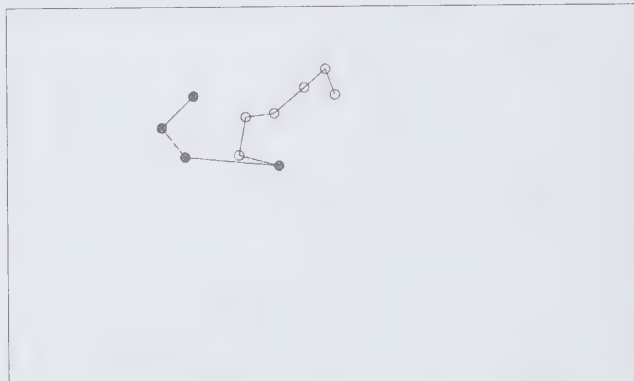


Figure 6 A random walk sample from the same population

limit random walk probabilities

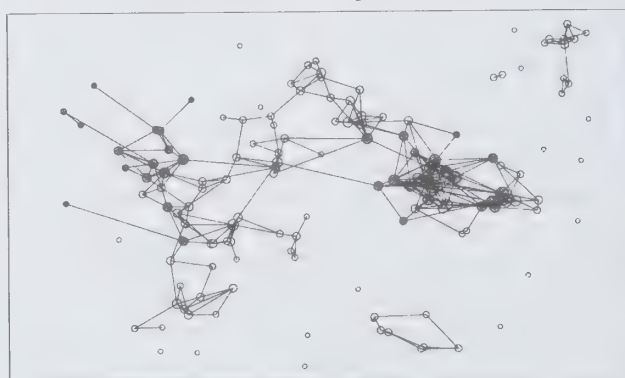


Figure 7 Random walk limit selection probabilities

2.1 Targeted random walk designs

One of the early motivations for using random walk designs with hidden populations was to penetrate deeper into the population, that is, farther from the initial sample and thereby obtain a more representative sample of the population. When the probabilities of selecting a given person by such a method are calculated either step by step or in their stationary limit, they are not in general equal but depend on the link and degree structure. With the motivation first to find a method for selecting a sample through a network such that the stationary probabilities would be the same for each person or node, uniform and targeted random walk sampling designs were developed (Thompson 2006a). An additional motivation was to find a more flexible and adaptable way to sample through a network.

Since a random walk with replacement through a graph or network is a Markov chain, ideas of Markov chain Monte Carlo can be applied to produce a different Markov chain having desired stationary probabilities. At each step of the sampling the state of the chain is the current node added to the sample. The stationary probabilities of the chain correspond to the stationary selection probabilities for each person or node. With a targeted walk design the random walk design is tweaked at each step, based on out-degree of each node, to obtain a design with specified limiting selection probabilities.

Suppose that at some step in the sampling person i is the last person who has been added to the sample. Using a random walk procedure we randomly select one of the links out from that person, and that link leads to person j , who is now our tentative selection. A screening interview reveals that person j has more links out than person i , so that the conditional probability of going from i to j as we just did is larger than the conditional probability in the reverse direction, since the transition probabilities are related to the reciprocal of the number of links out. Therefore we calculate a probability less than one and accept person j into the sample only with that probability. If our tentative selection is not accepted we independently again choose a link out from person i . The probability of acceptance of the candidate link is based on the Hastings (1970) generalization of the Metropolis algorithm. The acceptance probability depends on the desired target selection probabilities, the number of links out from the current node and the candidate node, and the probability of going in either direction with a random jump if that is part of the design (Thompson 2006a).

Note that the method depends only on links out, which can usually be determined for sample members, whereas links in to sample individuals usually can not be determined. Therefore the method applies to directional as well as symmetric networks.

A uniform walk design is the special case in which the targeted stationary selection probabilities are all equal. A targeted random walk design could be used for example to obtain a sample from a hidden population in which an individual with a certain high-risk behavior would have selection probability twice that of an individual without the behavioral characteristic.

It is the sample of accepted people or nodes that has the desired stationary selection probabilities. If the tentative selections had been interviewed thoroughly also, not only the screening interview about out-degree, then in principle the estimates from the accepted sample could be improved using the Rao-Blackwell method (Casella and Robert 1996). That would involve calculation of the probabilities of getting the same data with different accept-reject results and in different orders of selection. With each of the different accept scenarios the estimate would be computed using the accepted set and each value weighted by the ordered selection and acceptance probabilities. In most cases there are too many combinations for exact calculation, and a more practical approach would be the Markov chain resampling method at the inference stage described in a later section of this paper. It is not clear that in practice it would be desirable to compute the improved estimators using the data since full interviews rather than screening interviews would be required for those not initially accepted, the computations for the improvement are potentially demanding, and the calculation depends on knowing the selection probabilities for the initial sample, which is not needed for the simple estimators.

With a targeted walk design in which the target stationary selection probability π_i of node i is proportional to c_i , an asymptotically consistent estimator, based on the limiting probabilities, is provided by the generalized ratio estimator

$$\hat{\mu} = \frac{\sum_{s_a} y_i / c_i}{\sum_{s_a} 1 / c_i}$$

where y_i is the value of the variable of interest for the i^{th} node and s_a is the sample of selected nodes. In this type of estimator the relative values of target probabilities need be specified since the proportionality constant cancels out.

Note that a straight Horvitz-Thompson or Hansen-Hurwitz estimator can not be used because the proportionality constant in the inclusion probabilities is unknown, whereas in the generalized ratio estimator it cancels out. Again the limiting probabilities on which the estimator is based hold exactly for the with-replacement design. For the without-replacement variation, the properties of the targeted strategies were fairly closely approximated by the with-replacement properties in the empirical comparisons (Thompson 2006a).

2.1.1 Designs using weighted links

Many studies of socially networked populations conceptualize the network as having nodes (people) and lines or arrows representing the links or relationships between people. The network is characterized by an incidence matrix of 0s and 1s indicating when there is a link from node (row) i to node (column) j . In many real situations, however, more than one type of link may be of interest and links may have different weights representing differing strengths of a relationship. For example, in studies of risk behaviors and interventions in relation to the HIV epidemic, two types of links of high interest are sexual relationships and drug injecting relationships. Other social relationships, such as friendships and living arrangements, may also be of interest to investigators and may be useful in finding members of the population. These types of relationships may have weights corresponding to frequency of encounters, geographic proximity, or other measures of strength.

In the basic form of weighted link designs we consider, in which one link from the most recently selected person is selected from the links out from that person, the selection is made with probability proportional to link weight. More generally, the selection could be made based on that weight but not necessarily proportional to it. However, we could then redefine the weight to be proportional to the probability we have under the design of following that weight, so that the following result would still apply.

The following derivation shows that under suitable conditions the stationary selection probability for each person with such a design is proportional to the sum of the link weights out from that person. The result applies for a population in which it is possible to reach any one person from another following some path in which each link has weight greater than zero. That is, the population has a single component.

For such a condition to hold it is advantageous to have at least some probability of following common but weak links. For example, a study of a sexually transmissible epidemic may want to focus with high probability on sexual links. But sexual links do not connect the population into a single component. Therefore, some smaller probability is allowed in the design for following friendship or geographic links, which represent weaker relationships between people and are of less inherent interest to investigators but serve to connect the population. Thus, the combination of different types of links in this situation turn the population into a single component for purposes of the design.

2.1.2 Stationary distribution of weighted link Markov chain design

In this section we derive the stationary distribution of a weighted link design in a single component situation. Keep

in mind that we may create the single component property through innovative use of geographic links in combination with social links.

Let w_{ij} be the weight of a link between node i and node j , and assume that these links are symmetric, so that $w_{ij} = w_{ji}$. Consider a random walk design, with replacement, in which the transition probability to node j , given the walk is at node i , is proportional to w_{ij} . That is, one link is selected out from node i with probability proportional to weight. The transition probability is thus $P_{ij} = w_{ij} / w_{i\cdot}$. The sum $w_{i\cdot} = \sum_j w_{ij}$ is the total weight out from node i , generalizing the concept of degree with equally weighted nodes.

Suppose the graph has only a single component, that is, any node in the graph can be reached from any other node by a path in which every link has positive weight. Then the stationary probability for node i is proportional to $w_{i\cdot}$.

Suppose that the probability that the walk is at node i at time t is $\pi_i = w_{i\cdot} / w_{\cdot\cdot}$, for $i = 1, \dots, N$, where $w_{\cdot\cdot} = \sum_i \sum_j w_{ij}$, the total of all the weights. Then the probability that the process is at node i at time $t + 1$ is $\sum_j \pi_j P_{ji}$ by the law of total probability. In terms of the link weights, this sum is $\sum_j (w_{j\cdot} / w_{\cdot\cdot}) (w_{ji} / w_{j\cdot}) = \sum_j w_{ji} / w_{\cdot\cdot}$. Because of the symmetry of the weighted links, this becomes $w_{i\cdot} / w_{\cdot\cdot}$, so that if node i has this probability at time t it has the same probability at time $t + 1$, so that these are the stationary probabilities of the process. By induction, once the process reaches its stationary distribution it remains in it for every step thereafter. In practice, especially with small sample sizes or with different design variations, the stationary distribution serves as an approximation to the exact distribution.

If the weights are not symmetric, the selection probabilities of the random walk design will still approach a stationary distribution provided there is only a single component or, if not, that the design incorporates random jumps. However, with the directional weighted links, the stationary distribution is no longer of the simple form that can be calculated from sample data.

2.1.3 Different uses of weighted link designs

Variations of weighted link designs could prove useful in situations of the following types.

- (1) Designs using general weights of links, on a continuous or discrete scale, representing strength or importance of relationships and probability of following them.
- (2) Situations with two types of links, represented by two weights, such as social networks with strong and weak relationship links, or an HIV-at-risk study focusing on both sexual contacts and drug using relationships.

- (3) Survey settings in which links represent the geographic or “random jump” part of the design, or the seed design. For example, all people within a given geographic stratum are linked by a geographic link, or all the people who visit any of the venues on an ethnographic map are thereby linked.
- (4) In a situation where a sampling frame exists but the frame covers only part of the population, all units within the frame can be considered to be connected by a “frame link”. Venue-based sampling typically forms one example of this type of situation.
- (5) Using a variation on the sampling design as a model for the way a virus or other infectious agent “samples” people in a population. A type of weighted link design could be developed as a model for the spread of an infectious disease, finding the different importance of different links. For influenza, the relative importance of air transported droplets (sneezing, coughing) versus indirect contact through solid objects (door knobs, money). For HIV, the relative importance of different types of sexual contacts and unsafe injections, whether for illegal drugs or unsanitary medical injections especially in third world countries. The disease transmission in a simulation has a slightly different protocol than the implemented designs, in that instead of thinking of one new link selected at each selection time step, there could be anywhere from zero to a high number of transmissions in a time step.

2.1.4 Properties of weighted link designs and associated population graphs

Suppose the relationships in the population are assigned weights, with the weight w_{ij} denoting the strength of the relationships from node i to node j . And suppose we use a link tracing design of the walk type in which the transition probability is

$$P_{ij} = \frac{w_{ij}}{w_{i.}}$$

where $w_{i.} = \sum_{j=1}^N w_{ij}$. This is the conditional probability of selecting node j as the next sample unit, given the most recently selected unit is node i . The walk design is a Markov chain on a graph, in which the graph has weighted links.

We will next consider the question in the other direction of when a Markov chain can be represented by a design of this sort on a graph with weighted links. Given a Markov chain specified by a matrix of transition probabilities P_{ij} , we can always represent it as a walk design of this type on a graph with weighted links so long as the links satisfy the first of the following properties:

- (1) $w_{ij} = P_{ij} w_{i.}$, where the row weight totals are arbitrarily chosen.

Next consider imposing some property on the weight row totals to make them unique. For example:

- (2a) If the $w_{i.}$ weight row totals are chosen to be all equal to a constant such as one, then the link weights represent the conditional transition probabilities given the process is at the node at which they originate.
- (2b) If the $w_{i.}$ weight row totals are proportional to the stationary probabilities π_i of the Markov chain for each node i , or equal to them, then the weights represent “flows” of the Markov chain, that is, the unconditional probabilities of transitions along the links:

$$w_{ij} = P_{ij} \pi_i.$$

In the practical situations for which we are trying to find appropriate models and designs, the weights may be at least partially given by the natural circumstances of the situation. For example the weight w_{ij} may represent the presence or absence of a link from person i to person j , or the number of transactions of a certain type in a given time period from i to j . In that case, condition (2a) above would not in general be satisfied and condition (2b) would be satisfied only if all the weights were symmetric, that is, if $w_{ij} = w_{ji}$ for all i and j .

In particular, if some or all of the weights are asymmetric, with $w_{ij} \neq w_{ji}$, then (2a) would not usually be satisfied and it would not be possible to arbitrarily choose weights to impose the condition because typically the stationary probabilities would not be known and could not be calculated from the sample data. However, although the row totals $w_{i.}$ could not be arbitrarily imposed, they can be known for units in the sample since they are simply the total weight out from each unit.

2.2 Adaptive web sampling

Targeted random walk designs provide considerable flexibility and control not offered by regular random walks. The use of weighted links with these designs extends that flexibility farther. This flexibility is still constrained, however, by the restriction that the selection of the next link to follow can depend only on the most recently selected node in the sample. The incentive for developing the next set of designs was to remove this restriction and greatly expand the scope for flexibility and control in the available strategies.

In an adaptive web sampling design (Thompson 2006b) an initial sample of one or more unit/node is selected by simple random sampling or other conventional design. From then on, at each step in the sampling there is an active set consisting of the sample selected so far or some subset of it.

In the simplest case, one link is selected from the links out from this set. Sampling continues in this fashion until the desired sample size or some other stopping criteria has been satisfied. Some small probability is allowed, however, that the next node is selected at random, or by some other conventional design, from the entire population. The designs can be done with or without replacement.

More generally a set of links can be selected at each step. Also the links at each step can be selected by a design more complicated than simple random sampling. The selection probabilities can be dependent on node or link characteristics and can be varying over time.

The basic idea of an adaptive web sampling design is shown in the next set of figures. In Figure 8, an initial sample of two nodes has been selected by random sampling without replacement. At the next step a link may be chosen out at random from either of the initial nodes to add a new node to the sample, as shown in Figure 9. The next node is selected by following one of the links out from the current sample. With a random walk a link would need to be followed from the last node selected, but with adaptive web sampling any eligible link out from the current sample (active set) may be followed. Note the next selection, shown in Figure 10, is not via a link from the most recently selected node, but from a previous one. As sampling progresses it is free to branch out flexibly in different directions as well as select new nodes at random from the population (Figure 11). The design can be stopped at a specified sample size or some other criteria. In the design shown in the figures, links out from the current sample were not selected completely at random but with higher probability given to following links from high-risk individuals, represented by dark or red nodes. Further, the design shown allowed a 0.1 probability of selecting the new node at random at any step instead of following a link.



Figure 8 The first two nodes selected at random

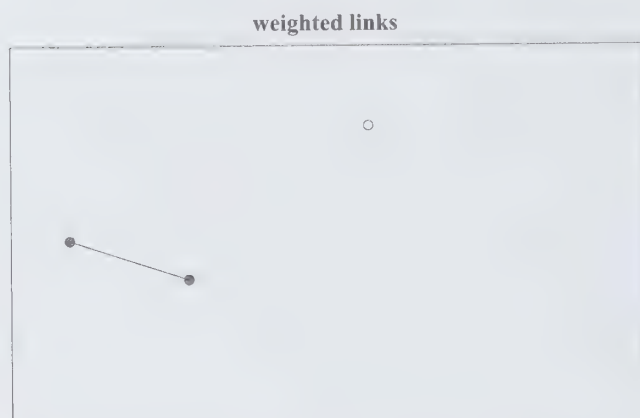


Figure 9 The next node is selected by following one of the links out from the current sample

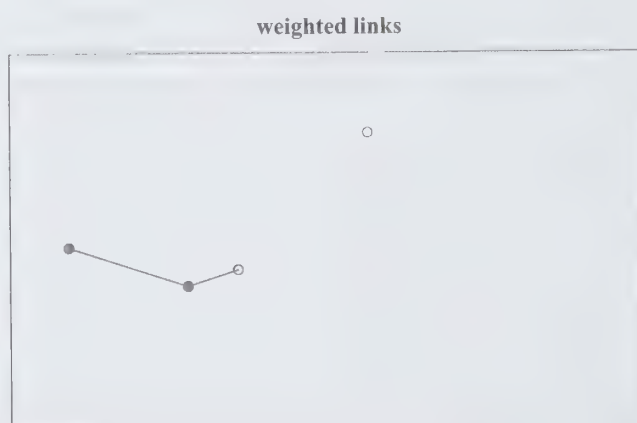


Figure 10 Note the next selection is not via a link from the last-selected node, but from a previous one

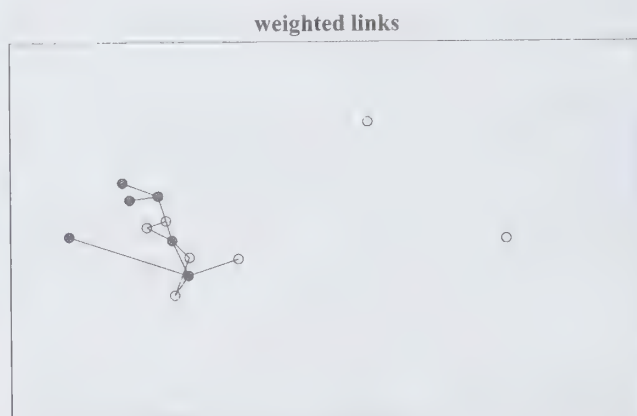


Figure 11 As sampling progresses it is free to branch out flexibly in different directions as well as select new nodes at random from the population

2.2.1 Inference methods

Design-unbiased and design-consistent estimation methods for use with adaptive web sampling designs are described in Thompson (2006b). Bayes model-based estimation methods for use with adaptive web sampling are described in Kwanasai (2005).

The design-based estimators are constructed by starting with some relatively easy to compute estimator that depends on the order of selection of the sample. This initial estimator is then improved using the Rao-Blackwell method, that is by obtaining the expected value of the initial estimator conditional on the minimal sufficient statistic.

2.3 Estimator based on initial sample mean

Suppose $\hat{\mu}_0$ is an unbiased estimator of the population mean that depends on the order in which the sample is selected. If the initial sample of nodes has been selected by simple random sampling, one example of an unbiased initial estimator that depends on order is the initial sample mean. The improved estimator has the form

$$\hat{\mu} = E(\hat{\mu}_0 | d_r) = \sum_{\{s: r(s)=s\}} \hat{\mu}_0(s) p(s | d_r).$$

Here s denotes the sample in order of selection, r is the reduction function that reduces the ordered sample to s , the unordered sample of the minimal sufficient statistic. The reduced data d_r consists of the unordered sample together with the associated values of the variables of interest. The improved estimator $\hat{\mu}$ is the expected value of the initial estimator over all $n!$ reorderings of the sample data. In calculating the expectation, each of the reorderings is weighted by the selection probability $p(s | d_r)$.

Other initial estimators used with adaptive web sampling utilize the entire sample data but depend on order and are based on using the conditional probabilities of selecting each new unit in sequence given the previously selected units. Four types of design-based estimators for use with adaptive web sampling are given in Thompson (2006b).

Computation of the improved estimator $\hat{\mu}$ and its variance estimators under various adaptive web designs involves enumerating the reorderings of the sample selection sequence. For each reordering, the probability of that ordering under the design is computed, along with the values of the estimators and variance estimators. Direct calculation is fast and efficient up to sample sizes of ten or so, which involve no more than a few million permutations to be enumerated. For larger sample sizes, the numbers of permutations or combinations of potential selection sequences in the conditional sample space become prohibitively large for the exact, enumerative calculation. For this reason, a Markov chain resampling approach was used in Thompson (2006b) for computing the improved estimators.

The resampling procedure is as follows. The object is to obtain a Markov chain x_0, x_1, x_2, \dots having stationary distribution $p(x | d_r)$. Here x_k denotes an entire reordering of the sample at step k of the chain. Suppose that at step $k-1$ the value is $x_{k-1} = j$, so that h denotes the current permutation of the sample data in the chain. A tentative or

candidate permutation c_k is produced by applying the original sampling design, with sample size n , to the data as if the sample comprised the whole population, that is, as if $N = n$. This resampling distribution, denoted p_c differs from, but has some similarity to, the actual sampling design p . The desired conditional distribution $p(x | d_r)$ is proportional to the unconditional distribution $p(x)$ under the original design applied to the whole population.

Let

$$\alpha = \min \left\{ \frac{p(c_k) p_c(x_{k-1})}{p(x_{k-1}) p_c(c_k)}, 1 \right\}.$$

With probability α , t_k is accepted and $x_k = c_k$, while with probability $1 - \alpha$, c_k is rejected and $x_k = x_{k-1}$.

This procedure produces a Markov chain x_0, x_1, x_2, \dots having the desired stationary distribution $p(x | d_r)$. The chain is started with the original sample s in the order actually selected. Given any value of the minimal sufficient statistic d_r , the chain is thus started in its stationary distribution and so remains in its stationary distribution step by step.

Suppose that n_r resampled permutations are selected by this process and let $\hat{\mu}_{0h}$ denote the value of the initial estimator for the h^{th} permutation. An enumerative estimator of the form $\hat{\mu} = E(\hat{\mu}_0 | d_r)$ is replaced by the resampling estimator

$$\tilde{\mu} = \frac{1}{n_r} \sum_{h=0}^{n_r-1} \hat{\mu}_{0h}.$$

Bayes model-based inference with adaptive web sampling designs also requires the use of Markov chain Monte Carlo (MCMC) methods except in certain fairly simple design situations (Chow and Thompson 2003) where explicit Bayes posterior distribution, estimators, and intervals can be obtained. More generally the MCMC sequence involves at each step updating of model parameter estimates and, in a data augmentation procedure, obtaining a complete realization of the population network and its values from the predictive posterior distribution conditional on the observed data (Kwanisai 2005, 2006). The resulting Markov chain sequence of complete population realizations provides the flexibility to make inference about many types of population characteristics.

2.4 Modification of adaptive web sampling procedures

Adaptive web sampling designs are a generalization of random walk designs. The more general designs do not have the exact stationary distribution properties of walk designs, since more than one link may be followed from any node, links may be followed from sample nodes other than the most recently selected one, and the sampling may be done

without replacement. However, the stationary distribution properties of a random walk or other Markov chain design may serve as a guide to approximate properties one might expect from a similar adaptive web sampling design.

During the sampling, at the time of the t^{th} unit selection in the k^{th} wave, let $w_{a_{kt}+}$ be the total number of links out, or the total of the weight values, from the active set a_k to units not in the current sample s_{ckt} . That is, $w_{a_{kt}+} = \sum_{\{i \in a_k, j \in \bar{s}_{ckt}\}} w_{ij}$. When w is an indicator variable, $w_{a_{kt}+}$ is the total of the net out-degrees of the individual units in the active set a_k , where net out-degree is the out-degree of a unit minus the number of its links to other units already in the current sample.

For each unit i in the sample, the variable of interest y_i and the out-degree (or out-weight) w_{i+} are recorded. In addition, for each pair of units (i, j) for which both i and j are in the sample, the values of the link variables w_{ij} and w_{ji} are observed.

Consider as a candidate for the t^{th} selection in the k^{th} wave a unit i not in the current sample, so $i \notin s_{ckt}$. Suppose the current active set a_k contains one or more units having links or positive weights out to unit i , and let $w_{a_k i} = \sum_{j \in a_k} w_{ij}$ denote their total. The probability that unit i is the next unit selected is

$$q_{kti} = b \frac{w_{a_k i}}{w_{a_{kt}+}} + (1 - b) \frac{1}{(N - n_{s_{ckt}})}$$

where b is between 0 and 1. If there are no links at all out from the current active set, then

$$q_{kti} = \frac{1}{(N - n_{s_{ckt}})}.$$

Thus, with probability b link-tracing is done, and one of the links out from the current active set is selected at random, or with probability proportional to its weight, and the node to which it leads is added to the sample, while with probability $1 - b$ the new sample unit is selected completely at random from the units not already selected. However, if there are no links or positive weights out from the active set to any unsampled units, then the next unit is selected from the collection of unsampled units.

Basic adaptive web sampling can be generalized to use weighted links. If the relationship variable w consists of weights, instead of having just 0 or 1 values, then the link-based selection can depend on these weights. For example, link weights can be defined in relation to the y value of an originating node or as a distance measure to the connected node, so that links are followed with higher probability from nodes with higher values or with lower probability to distant nodes. Then a link from the active set can be selected with probability proportional to link weight, or with some other selection probability $p(i | s_{ckt}, a_k, y_{a_k}, w_{a_k})$ depending on

variables of interest only through the active set. For example, a link out could be selected at random from the links with w_{ij} greater than some constant, or y_i greater than some constant. The selection probability when links are not followed does not have to be uniform over the units not in the current sample, but can be a more general design $p(i | s_{ckt})$ such as selecting with probability related to an auxiliary variable or from a spatially defined distribution.

With weighted links w represents a possibly continuous link weight variable and the probability that unit i is the next unit selected is

$$q_{kti} = bp(i | s_{ckt}, a_k, y_{a_k}, w_{a_k}) + (1 - b) p(i | s_{ckt}).$$

If there are no links or positive weights from a_k to i , then

$$q_{kti} = p(i | s_{ckt}).$$

Once unit i has been selected, it is possible to add an accept/reject step for deciding whether to include it in the active set, for example, accepting with higher probability if unit i has a high value or high degree.

In the design the constant b itself can also be replaced by a probability $b(k, t, a_k, y_{a_k}, w_{a_k})$ depending on values related to nodes and links in the active set or changing as sample selection progresses. For example, if the values of the units in a_k are particularly high, we could increase the probability of following links. As for dependence of b on (k, t) , the use of an initial conventional sample of size $n_0 > 1$ may be viewed as serving to obtain some information from basic coverage of the population before adaptive sampling is allowed to commence.

3. Spatial adaptive web sampling

Adaptive sampling designs such as adaptive cluster sampling (Thompson 1990) were developed in response to the need for more effective strategies for sampling spatially uneven populations, particularly those having a rare, clustered geographic distribution. Most populations having a network structure also have an inherent geographic or spatial structure. For example, human populations have social network structure but are also distributed in space. Of particular interest from the sampling design point of view, spatial structures can be characterized with graph or network structures. For example, neighborhood relationships based on geographic proximity can be recast in the form of lattice-type graphs. In this way, network designs such as those described in the previous section can be applied to solve spatial sampling problems.

In this section the use of adaptive web sampling designs to sample a spatially uneven population will be described.

These designs could be viewed as a generalization of adaptive cluster sampling. In this view, adaptive cluster sampling would be a special case in which every link is followed until there are no more links out from the current sample. The adaptive web sampling class of designs offers more flexibility and control, however, and is potentially more efficient to use for many spatial populations.

With adaptive cluster sampling the constraint to continue to sample until all neighbors of all units satisfying the condition were included meant that overall sample size was not controlled in advance and was rather stringent when some networks were unusually large. Adaptive web sampling in the spatial context solves this problem since sample size can be fixed in advance. In terms of its network recasting, the simple unbiased estimators of adaptive cluster sampling use data only from the strongly connected components that the initial sample intersects. Rao-Blackwell improvements based on those estimators can use in addition data from the weakly connected extensions of those components. The familiar edge units of spatial adaptive cluster sampling are a special case of such weakly connected extensions of strongly connected components.

Figure 12 depicts a study region with a spatial clustered population as may be encountered in ecological, epidemiological, and social demographic surveys. In one form of adaptive spatial designs the neighborhood of a unit is defined as the set of immediately adjacent units, and neighboring units are added to the sample when the value of a sample unit is high or meets some other criterion. In Figure 13 the spatial population has been recast as a directed graph. The square spatial units are redrawn as nodes in a graph, and whenever the number of objects in a unit exceeds zero, arrows representing graph links are drawn from that node to neighboring nodes. Nodes representing units with nonzero values are colored dark (red). Figure 14 shows a random sample of nodes to be used as the initial sample of an adaptive web design. The adaptive web sampling continues until the targeted final sample size of 20 units is obtained in Figure 15. The sample is recast in the spatial setting in Figure 16. Unlike adaptive cluster sampling, it was not necessary to continue sampling until every unit in a sampled connected component is included. Further, the small probability of a random jump keeps the design from being stuck in any connected component.

A glimpse of the immense flexibility offered with the adaptive web sampling designs in the spatial setting is shown in Figure 17. In the top row a spatial population is recast as a graph, though the directions of the links are not shown. The bottom row shows samples from two variations of adaptive web sampling. On the left, sixteen initial units have been selected independently at random. From each, an adaptive web sampling procedure is carried out to a sample

size of five units. With this design, the sample is spread throughout the study region while also reaching into components. In the design on the right a single initial unit is selected at random and adaptive web sampling continues to a total of 80 units. The 0.1 probability of selecting the next unit at random at any step prevents the design from being stuck in any one component. With this design the main components or aggregations get very thorough, though not exhaustive, coverage.

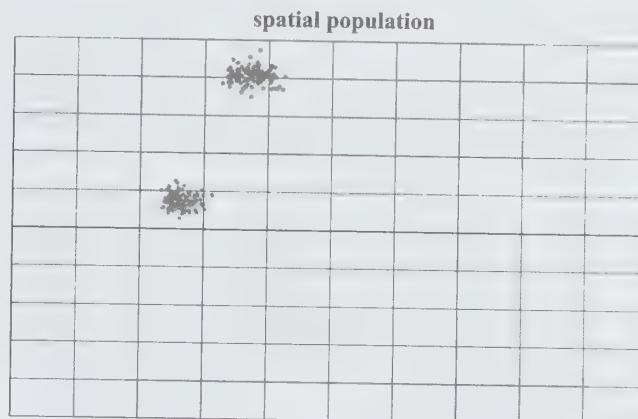


Figure 12 A spatially clustered population

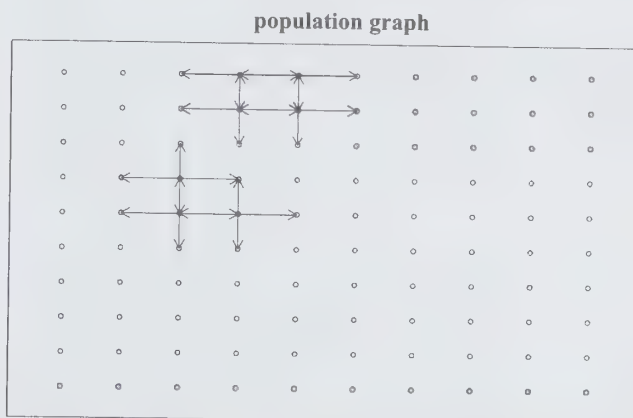


Figure 13 A network representation of relevant neighborhood relationships in the spatial population



Figure 14 An initial random sample of spatial units

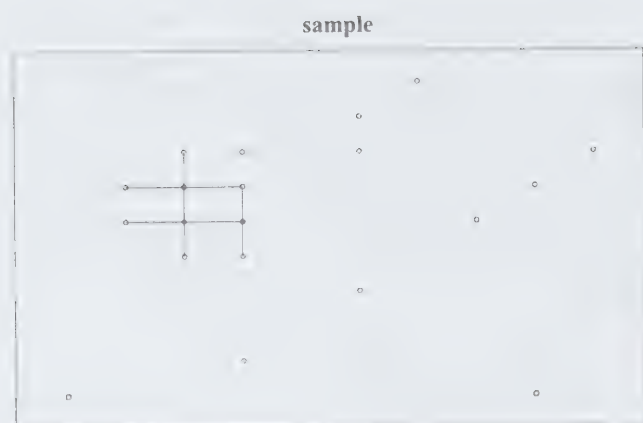


Figure 15 Adaptive web sample of 20 units starting from the initial sample of the previous figure

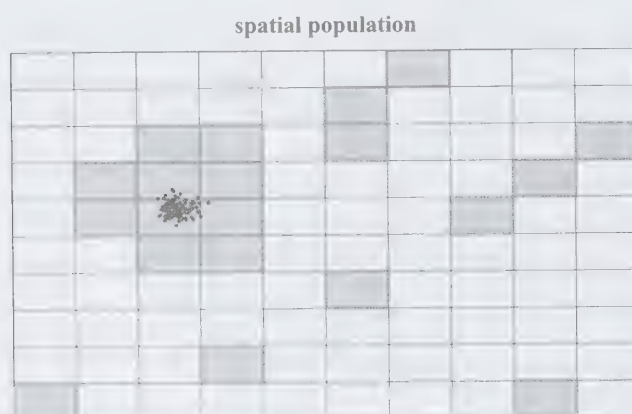


Figure 16 Spatial representation of the adaptive web sample

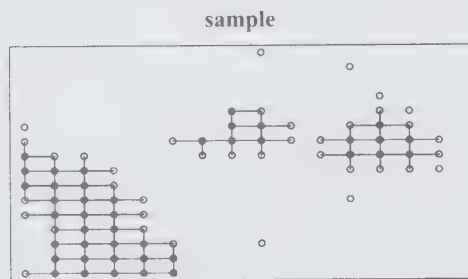
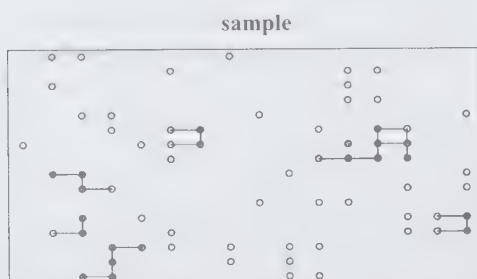
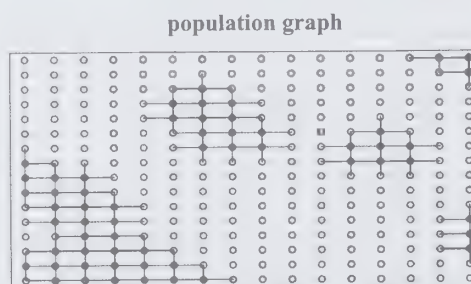
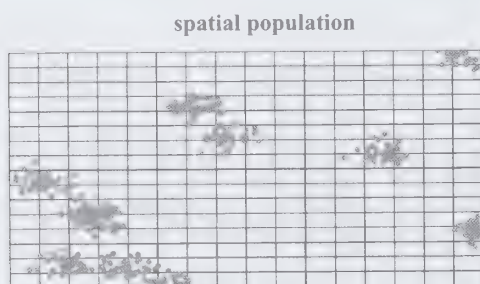


Figure 17 Adaptive web sampling design variations

3.1 Spatial designs with weighted links

For selecting spatial samples, link weights can be defined as a function of the distance between sites. For example, for increased sample the function would give larger weight to sites at close distance. On the other hand, for space filling purposes sites at larger distance could have larger weight. A network sampling design in such a setting, with link weights defined solely on the basis of distance, would not in general be adaptive. That is because the spatial frame would enable a link-tracing design to select the entire sample of sites before going in the field to make any observations.

More generally though link weights can be defined as a function of both weights and observed values. For a unit in the sample having a high observed value of the variable of interest, the function could give higher weight at distances close to that site and smaller weight to distance sites. For a unit having a low value of the variable of interest the weight function could have a more uniform shape.

Random walk designs in particular are straightforward to carry out in spatial settings with links weights dependent only on distance. That is because at any point in the sampling the selection of the next site depends only on the most recently selected site, so that only one weight function needs to be considered. With more general designs such as adaptive web sampling the use of link weight functions dependent on both distance and value opens up very wide flexibility in the possibilities available for adaptive strategies.

4. Discussion

Adaptive sampling designs expand considerably the possibilities for sampling strategies. They appear to be especially useful for populations which are otherwise difficult to sample. Network sampling designs are inherently adaptive in most cases and can provide more effective ways to sample populations with network or spatial structure. In this paper the emphasis has been on designs obtaining low mean square error or providing practical means of reaching a hidden population. In other cases the primary objective might be simply to obtain a higher yield sample, that is, a sample having a high total value of the variable of interest. For instance environmental hot spots is where remediation must be made, high risk components of a epidemic related network where treatment or intervention might have the greatest effect. The advantages of an adaptive approach are even more straightforward when the objective is high sample yield.

Fully optimal sampling strategies are in most cases not practical to implement, because of computational complexity and model dependency. A more practical approach is to make improvements over conventional designs with simple adaptive procedures that capture much of the essence, and the choice of design often having much more effect than one inference method versus another.

Simulation analyses with adaptive strategies of different types have tended to lend support to the idea that it is good to have a strong underlying conventional component. Many of the practical strategies have the form of an initial conventional sample with adaptive sampling extending the sample from there through either network or spatial relationships and depending on observed values. Strategies with that type of balance between conventional and adaptive components have in simulations generally performed better than, say, selecting a single unit conventionally and adaptively adding the whole rest of the sample from there. In the simulations most efficient strategies tended to have an initial sampling making up about 60-80 percent of the total sample size. The modest amount of adaptive sampling after that then produced large gains in efficiency. This empirical experience goes along with the characteristic of optimal adaptive strategies, in which there seems to be a push and pull between spreading units far apart or filling in unobserved parts of the study region, corresponding to the conventional component of the simplified designs, and placing new units in the most promising areas, corresponding to the adaptive component in the simplified designs.

Acknowledgements

The author would like to thank the Census Bureau and in particular Patrick Flanagan for hosting a lecture series on

which this paper is in part based. Research on the sampling designs described has been supported by the Natural Science and Engineering Research Council, the National Science Foundation, the National Center for Health Statistics, Centers for Disease Control and Prevention, Los Alamos National Laboratories, and the National Institutes of Health.

References

- Basu, D. (1969). Role of the sufficiency and likelihood principles in sample survey theory. *Sankhyā*, A, 31, 441-454.
- Birnbaum, Z.W., and Sirken, M.G. (1965). Design of sample surveys to estimate the prevalence of rare diseases: Three unbiased estimates. *Vital and Health Statistics*, Series 2, No. 11. Washington: Government Printing Office.
- Brin, S., and Page, L. (1998). The anatomy of a large-scale hypertextual web search engine. *Proceedings of the 7th International World Wide Web Conference*. Elsevier, 107-117.
- Casella, G., and Robert, C.P. (1996). Rao-Blackwellization of sampling schemes. *Biometrika*, 83.
- Chao, C.-T., and Thompson, S.K. (2001). Optimal adaptive selection of sampling sites. *Environmetrics*, 12, 517-538.
- Chow, M., and Thompson, S.K. (2003). Estimation with link-tracing sampling designs - A Bayesian approach. *Survey Methodology*, 20, 197-205.
- Felix-Medina, M.H., and Thompson, S.K. (2004). Combining link-tracing sampling and cluster sampling to estimate the size of hidden populations. *Journal of Official Statistics*, 20, 19-38.
- Frank, O. (1971). *Statistical Inference in Graphs*. Stockholm: Försvarets Forskningsanstalt.
- Frank, O. (1977a). Survey sampling in graphs. *Journal of Statistical Planning and Inference*, 1, 235-264.
- Frank, O. (1977b). Estimation of graph totals. *Scandinavian Journal of Statistics*, 4, 81-89.
- Frank, O. (1978a). Estimating the number of connected components in a graph by using a sampled subgraph. *Scandinavian Journal of Statistics*, 5, 177-188.
- Frank, O. (1978b). Sampling and estimation in large social networks. *Social Networks*, 1, 91-101.
- Frank, O. (1979). Estimation of population totals by use of snowball samples. In *Perspectives on Social Network Research*, (Eds., P.W. Holland and S. Leinhardt). New York: Academic Press, 319-347.
- Frank, O., and Snijders, T. (1994). Estimating the size of hidden populations using snowball sampling. *Journal of Official Statistics*, 10, 53-67.
- Hastings, W.K. (1970). Monte-Carlo sampling methods using Markov chains and their application. *Biometrika*, 57, 97-109.
- Hekathorn, D.D. (1997). Respondent driven sampling: A new approach to the study of hidden populations. *Social Problems*, 44, 174-199.
- Hekathorn, D.D. (2002). Respondent driven sampling II: Deriving valid population estimates from chain-referral samples of hidden populations. *Social Problems*, 49, 11-34.

- Klov Dahl, A.S. (1989). Urban social networks: Some methodological problems and possibilities. In *The Small World* (Ed., M. Kochen). Norwood, NJ: Ablex Publishing, 176-210.
- Kwanisai, M. (2005). Estimation in Link-Tracing Designs with Subsampling. Ph.D. thesis. The Pennsylvania State University, University Park, PA, U.S.A.
- Kwanisai, M. (2006). Estimation in networked populations. *Proceedings of the Survey Research Section*, American Statistical Association, Washington, DC., 3285-3291.
- Lovász, L. (1993). Random walks on graphs: A survey. In *Combinatorics, Paul Erdős is Eighty* (Eds., D. Miklós, D. Sós and T. Szőni). János Bolyai Mathematical Society, Keszthely, Hungary, 2, 1-46.
- Salganik, M.J., and Heckathorn, D.D. (2004). Sampling and estimation in hidden populations using respondent-driven sampling. *Sociological Methodology*, 34, 193-239.
- Sirken, M.G. (1970). Household surveys with multiplicity. *Journal of the American Statistical Association*, 63, 257-266.
- Sirken, M.G. (1972a). Stratified sample surveys with multiplicity. *Journal of the American Statistical Association*, 67, 224-227.
- Sirken, M.G. (1972b). Variance components of multiplicity estimators. *Biometrics*, 28, 869-873.
- Thompson, S.K. (1990). Adaptive cluster sampling. *Journal of the American Statistical Association*, 85, 1050-1059.
- Thompson, S.K. (2002). *Sampling, Second Edition*. New York: John Wiley & Sons, Inc.
- Thompson, S.K. (2006a). Targeted random walk designs. *Survey Methodology*, 32, 11-24.
- Thompson, S.K. (2006b). Adaptive web sampling. *Biometrics*, 62, 1224-1234.
- Thompson, S., and Frank, O. (2000). Model-based estimation with link-tracing sampling designs. *Survey Methodology*, 26, 87-98.
- Thompson, S.K., and Seber, G.A.F. (1996). *Adaptive Sampling*. New York: John Wiley & Sons, Inc.
- Zacks, S. (1969). Bayes sequential designs of fixed size samples from finite populations. *Journal of the American Statistical Association*, 64, 1342-1349.

Alternative survey sample designs: Sampling with multiple overlapping frames

Sharon L. Lohr¹

Abstract

Designs and estimators for the single frame surveys currently used by U.S. government agencies were developed in response to practical problems. Federal household surveys now face challenges of decreasing response rates and frame coverage, higher data collection costs, and increasing demand for small area statistics. Multiple frame surveys, in which independent samples are drawn from separate frames, can be used to help meet some of these challenges. Examples include combining a list frame with an area frame or using two frames to sample landline telephone households and cellular telephone households. We review point estimators and weight adjustments that can be used to analyze multiple frame surveys with standard survey software, and summarize construction of replicate weights for variance estimation. Because of their increased complexity, multiple frame surveys face some challenges not found in single frame surveys. We investigate misclassification bias in multiple frame surveys, and propose a method for correcting for this bias when misclassification probabilities are known. Finally, we discuss research that is needed on nonsampling errors with multiple frame surveys.

Key Words: Bias correction; Dual frame survey; Misclassification; Mode effects; Sampling for rare events; Sampling weights; Small area estimation.

1. Uses of multiple frame surveys

In classical design-based sampling theory, a probability sample is taken from the (single) sampling frame, and the inclusion probabilities in the sampling design can be used to make inferences about the population. Let y_i be a measurement on unit i in the population of N units, let S denote the set of units in the sample, and let $\pi_i = P$ (unit i is included in the sample). Then the Horvitz-Thompson (1952) estimator of the population total $Y = \sum_{i=1}^N y_i$ is $\hat{Y} = \sum_{i \in S} w_i y_i$, where $w_i = 1 / \pi_i$ is the sampling weight. If the sampling frame includes everyone in the target population, all sampled units respond, and there is no measurement error, then the Horvitz-Thompson estimator is unbiased for Y .

The practical challenges of sampling in the 1940s and 1950s drove the methodological developments of stratified multistage surveys and estimators such as the Horvitz-Thompson estimator. In-person surveys relied on unequal probability sampling to balance interviewer workloads and reduce variances. Response rates were high in many government surveys so that the assumptions for the Horvitz-Thompson estimator were reasonable. We now face new challenges in household surveys. Nonresponse rates are increasing, which means that survey estimates rely more on models. The ethnic and language diversity of a population can result in undercoverage and measurement error. Increasing technological diversity means that different residents may be best reached by different sampling modes; one must then be confident that different sampling modes measure the same quantities. Costs of collecting data have risen greatly, in part due to increasing nonresponse; at the

same time, governmental and research demands for data have also risen greatly.

Multiple frame surveys can achieve better population coverage at lower cost. They can be used as part of a structure of modular survey design that relies on different sampling frames to help reduce costs and achieve better coverage. They can also use administrative data efficiently. In this paper, we describe different types of multiple frame surveys and discuss some of the research that is completed and research that may be needed for their use.

One of the earliest multiple frame surveys (aside from early capture-recapture methods) was performed by the Census Bureau in 1949 (Hansen, Hurwitz and Madow 1953). In the Sample Survey of Retail Stores, a probability sample of primary sampling units (psus) was chosen. Within each psu, a list of large retail firms was constructed from records of the Old Age and Survivors Insurance Bureau. All firms on the list were sampled, and an area sample of firms in the psu that were not on the list was taken. In this case, a *screening* dual frame design was employed within each selected psu; units in the list frame were screened out of the area frame before sampling. Thus, the estimator of total sales summed the two estimators within each psu. No new statistical methods were required to estimate total sales in this survey, since essentially a stratified sample was taken in each psu: the firms on the list in the psu formed one stratum, and the firms in the area frame but not on the list in the psu formed the second stratum. The survey resulted in cost savings because it was relatively inexpensive to sample from the firms on the list, yet full coverage was obtained by also using the area frame.

1. Sharon L. Lohr, School of Mathematical and Statistical Sciences, Arizona State University, Tempe AZ 85287-1804. E-mail: sharon.lohr@asu.edu.

Many agricultural surveys also have used a screening dual frame survey design (González-Villalobos and Wallace 1996). In such a design, farms belonging to the list frame are removed from the area frame before sampling commences. Considerable cost savings can be realized since often the list frame is much less expensive to sample and it contains the largest farms.

In many cases, however, it may not be possible or practical to remove list-frame units from the area frame before sampling. Instead, in an overlapping dual frame survey, independent probability samples are taken from frame A (the area frame) and frame B (the list frame); this is depicted in Figure 1. Rare populations can often be sampled more efficiently using a multiple frame sample (Kalton and Anderson 1986). In an epidemiology study, for example, frame A might be that used for a general population health survey, while frame B might be a list frame of clinics specializing in a certain disease. The sample from frame B is expected to yield a high percentage of persons with the disease of interest, so that sampling will be efficient; the sample from frame A, though more expensive, leads to complete coverage of the population.

In other situations, all frames are incomplete, as considered by Hartley (1962); for example, frame A in Figure 2 might be a frame of landline telephones and frame B might consist of cellular telephone numbers. There are three domains: domain *a* consists of units in frame A but not in frame B, domain *b* consists of units in frame B but not in frame A, and domain *ab* consists of units in both frames. In the telephone context, domain *a* contains individuals belonging to a landline-only household, domain *b* consists of individuals who have only a cellular telephone, and domain *ab* consists of individuals who have both cellular and landline telephones. It is unknown in advance whether a household member sampled using one frame also belongs to the other frame (Brick, Dipko, Presser, Tucker and Yuan 2006); typically, respondents are asked about their cellular and landline telephone usage to determine domain membership.

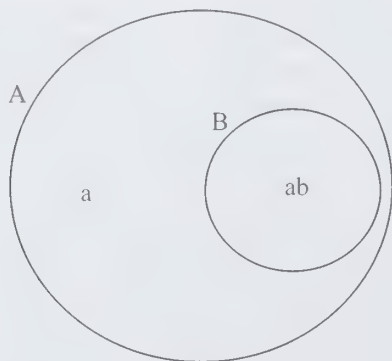


Figure 1 A dual frame design in which frame B is a subset of frame A

More than two frames can be employed as well, as illustrated in Figure 3 for a three-frame survey in which all frames are incomplete. In this situation, there are seven domains. Iachan and Dennis (1993) gave an example of a three-frame survey used to sample the homeless population, where frame A is a list of soup kitchens, frame B is a list of shelters, and frame C consists of street locations. Figure 4 displays a 3-frame survey in which frame A has complete coverage, while overlapping frames B and C are both incomplete but are less expensive to sample. This design has been used for the U.S. Scientists and Engineers Statistical Data System (SESTAT; National Science Foundation 2003) surveys. The same design might be used when A is the frame for a general population survey, B is a landline telephone survey, and C is a cell phone survey.

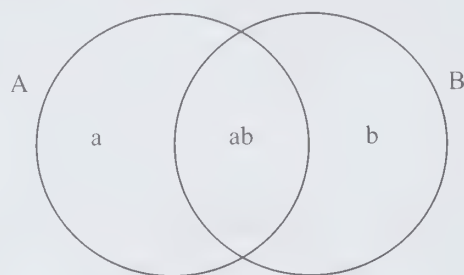


Figure 2 Frames A and B overlap, creating the three domains *a*, *b*, and *ab*

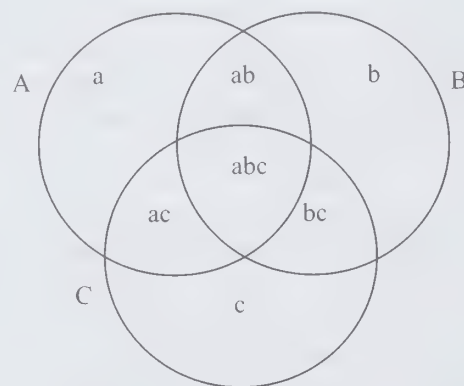


Figure 3 Frames A, B, and C are all incomplete and overlap

There is much potential for using multiple frame designs in household surveys, including:

1. Use of multiple list frames from administrative records.
2. Multiple mode sampling (for example, using independent samples from a cellular telephone frame and a landline telephone frame).

3. Future use of the internet for data collection. Although the internet presents many coverage and domain specification challenges, it is worthy of consideration because of the potential cost savings and ease of data collection and processing.
4. Improved small area estimation. A national survey is supplemented with smaller, localized surveys to obtain higher precision in those areas.
5. Improved estimation for rare populations. A general population survey may be supplemented by a survey from a frame with a high concentration of members of the rare population.
6. Modular survey design. A multiple frame approach can give more flexibility for design of continuing surveys. As particular frames become less expensive to sample, the relative allocation of sample size to the different frames can be modified. The modular approach also allows more flexibility in responding to changing needs for data.

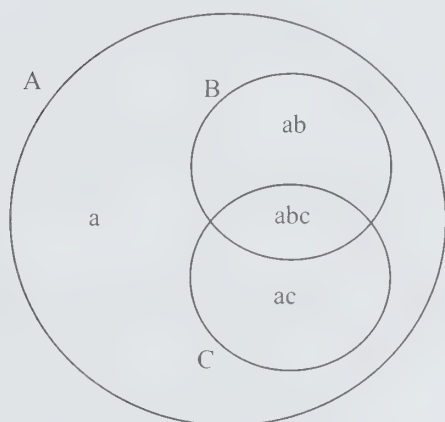


Figure 4 Frame A contains the entire population; frames B and C overlap and are both contained in frame A

The increased flexibility of multiple frame surveys comes at the cost of additional complexity, however. Information from the surveys must be combined to estimate population quantities, and there are many options for estimators. Section 2 summarizes estimators that have been developed for population totals and describes how these modify the sampling weights; Sections 3 and 4 discuss weight calibration and describe how to use survey software packages with multiple frame survey data. Nonsampling errors need to be considered in each frame singly, and in terms of their effect on estimates calculated from the combined information. Section 5 discusses effects of non-response and mode effects in multiple frame surveys.

In addition to the nonresponse, undercoverage, and measurement error problems that plague single frame surveys, multiple frame surveys may have domain misclassification.

The weight modifications for the estimators in Section 2 depend on the domain membership of the observations. If some observations in domain a are likely to be mistakenly recorded as belonging to domain ab , estimators may have substantial bias. We study effects of domain misclassification in Section 6, and propose a new method for adjusting for misclassification bias when misclassification probabilities are known. Finally, Section 7 discusses design issues and Section 8 discusses the potential and challenges of multiple frame surveys.

2. Estimators in overlapping multiple frame surveys

In this section we review estimators for the population total Y from overlapping multiple frame surveys, along with the weight modifications induced by these estimators. For simplicity of notation, we concentrate on dual frame surveys in Section 2.1, and outline extensions to multiple frame surveys in Section 2.2. In a dual frame survey, we can write

$$Y = Y_a + Y_{ab} + Y_b,$$

where Y_a is the total of the population units in domain a , Y_{ab} is the total of the population units in domain ab , and Y_b is the total of the population units in domain b . A special case is estimating the population size $N = N_a + N_{ab} + N_b$, as discussed in Haines and Pollock (1998). We discuss estimating population quantities other than totals and means, and using data from multiple frame surveys in other analyses, in Section 4.

We first set out some desirable properties for estimators from multiple frame surveys.

1. An estimator should be approximately unbiased for the corresponding finite population quantity.
2. Estimators should be internally consistent: that is, if \hat{Y}_1 estimates the number of female engineers in the population, \hat{Y}_2 estimates the number of male engineers in the population, and \hat{Y}_3 estimates the total number of engineers in the population, then we should have $\hat{Y}_1 + \hat{Y}_2 = \hat{Y}_3$. Internal consistency preserves multivariate relationships in the data. In practical terms, internal consistency requires that one set of weights be used for all estimates.
3. An estimator should be efficient, with low mean squared error.
4. An estimator should be of a form that can be calculated with standard survey software such as SUDAAN or SAS PROC SURVEYMEANS. This allows analysts to work with the data without having to write and test new software. In practical

terms, one data file is created from the multiple frame survey. The file includes one column of weights to be used for calculating point estimates, and it contains either variables describing the survey designs for formula-based variance estimation, or columns of replicate weights for replication-based variance estimation.

5. An estimator should, if possible, be robust to non-sampling errors that might occur with multiple frame surveys.

2.1 Estimators and weight adjustments for dual frame surveys

Consider the overlapping dual frame survey depicted in Figure 2, where domain ab is nonempty. A probability sample $S(A)$ of size n_A is drawn from the N_A units in frame A, and an independent probability sample $S(B)$ of size n_B is drawn from the N_B units in frame B. Unit i in sample $S(A)$ has probability of inclusion π_i^A and weight w_i^A , and unit i in sample $S(B)$ has probability of inclusion π_i^B and weight w_i^B . The weights may be the inverses of the inclusion probabilities, or they may be poststratified to agree with population counts; it is assumed that estimators of population totals are approximately unbiased.

Then $E[\sum_{i \in S(A)} w_i^A y_i] \approx Y_a + Y_{ab}$ and $E[\sum_{i \in S(B)} w_i^B y_i] \approx Y_b + Y_{ab}$. Consequently, an estimator that combines the observations from both surveys with the original weights, $\sum_{i \in S(A)} w_i^A y_i + \sum_{i \in S(B)} w_i^B y_i$, is biased for the population total Y . If the domain means differ, the corresponding estimator of the population mean may also be biased.

The various estimators for the population total Y that have been proposed in the literature modify the weights so that the estimators are approximately unbiased. The modified weights, shown below for the different estimators, are of the form $\tilde{w}_i^A = m_i^A w_i^A$ and $\tilde{w}_i^B = m_i^B w_i^B$. The population total is then estimated by

$$\hat{Y} = \sum_{i \in S(A)} \tilde{w}_i^A y_i + \sum_{i \in S(B)} \tilde{w}_i^B y_i \quad (1)$$

and the population mean \bar{Y} is estimated by $\hat{\bar{Y}} = \hat{Y} / \hat{N}$ where

$$\hat{N} = \sum_{i \in S(A)} \tilde{w}_i^A + \sum_{i \in S(B)} \tilde{w}_i^B.$$

The estimators will be approximately unbiased, then, if $m_i^A \approx 1$ for $i \in a$, $m_i^B \approx 1$ for $i \in b$, and $m_i^A + m_i^B \approx 1$ for $i \in ab$. All of the estimators reviewed in this section satisfy the criteria needed for approximate unbiasedness in the absence of nonsampling errors (see Lohr 2009).

Fixed weight adjustments. The simplest weight modification to preserve approximate unbiasedness, described by Hartley (1962), takes

$$m_{i,\theta}^A = \begin{cases} 1 & \text{if } i \in a \\ \theta & \text{if } i \in ab \end{cases}, \quad m_{i,\theta}^B = \begin{cases} 1 & \text{if } i \in b \\ 1 - \theta & \text{if } i \in ab \end{cases}, \quad (2)$$

where $\theta \in [0, 1]$. Using the modified weights $\tilde{w}_i^A = m_{i,\theta}^A w_i^A$ and $\tilde{w}_i^B = m_{i,\theta}^B w_i^B$ in (1), the resulting estimator $\hat{Y}(\theta)$ can also be expressed using the estimated domain totals $\hat{Y}_a^A = \sum_{i \in S(A), i \in a} w_i^A y_i$, $\hat{Y}_{ab}^A = \sum_{i \in S(A), i \in ab} w_i^A y_i$, $\hat{Y}_{ab}^B = \sum_{i \in S(B), i \in ab} w_i^B y_i$, and $\hat{Y}_b^B = \sum_{i \in S(B), i \in b} w_i^B y_i$. The estimator

$$\begin{aligned} \hat{Y}(\theta) &= \sum_{i \in S(A)} m_{i,\theta}^A w_i^A y_i + \sum_{i \in S(B)} m_{i,\theta}^B w_i^B y_i \\ &= \hat{Y}_a^A + \theta \hat{Y}_{ab}^A + (1 - \theta) \hat{Y}_{ab}^B + \hat{Y}_b^B \end{aligned} \quad (3)$$

thus estimates the domain total Y_{ab} by a weighted average of the frame A estimator, \hat{Y}_{ab}^A , and the frame B estimator, \hat{Y}_{ab}^B .

For a fixed value of θ , the estimator $\hat{Y}(\theta)$ gives internal consistency since the same set of adjusted weights is used for all variables. The estimator is simple to use and implement. The efficiency of the estimator depends on the value chosen for θ . Brick *et al.* (2006) used $\theta = 1/2$ in their study of a dual frame survey in which frame A was a landline telephone frame and frame B was a cellular telephone frame, and the value of $\theta = 1/2$ is frequently recommended (see, for example, Mecatti 2007). When $\theta = 0$ or 1, the data in the overlap domain from one of the samples are discarded and the survey becomes a screening dual frame survey.

Optimal estimators. Hartley (1962, 1974) proposed choosing θ in (3) so that the variance of $\hat{Y}(\theta)$ would be minimized. The optimizing value of θ is

$$\theta_H = \frac{V(\hat{Y}_{ab}^B) + \text{Cov}(\hat{Y}_b^B, \hat{Y}_{ab}^B) - \text{Cov}(\hat{Y}_a^A, \hat{Y}_{ab}^A)}{V(\hat{Y}_{ab}^A) + V(\hat{Y}_{ab}^B)}.$$

Since the variances and covariances are generally unknown, they must be estimated from the data, giving

$$\hat{\theta}_H = \frac{\hat{V}(\hat{Y}_{ab}^B) + \hat{\text{Cov}}(\hat{Y}_b^B, \hat{Y}_{ab}^B) - \hat{\text{Cov}}(\hat{Y}_a^A, \hat{Y}_{ab}^A)}{\hat{V}(\hat{Y}_{ab}^A) + \hat{V}(\hat{Y}_{ab}^B)}.$$

Skinner and Rao (1996) showed that Hartley's estimator can be calculated using adjusted weights. The weight modifications for Hartley's estimator $\hat{Y}(\hat{\theta}_H)$ are given by (2), substituting $\hat{\theta}_H$ for θ . Since $\hat{\theta}_H$ is consistent for θ_H , Hartley's estimator is asymptotically optimal among all estimators of the form $\hat{Y}_a^A + \hat{Y}_b^B + \theta \hat{Y}_{ab}^A + (1 - \theta) \hat{Y}_{ab}^B$. The modified weights $\tilde{w}_{i,H}^A$ and $\tilde{w}_{i,H}^B$ are functions of the variances and covariances of estimated domain totals, however. This has two consequences: (1) the modified weights are random variables, and their variability needs to be accounted for in standard errors of estimators, and (2) the optimal weight modifications will differ for different response variables, leading to internal inconsistency.

Fuller and Burmeister (1972) proposed modifying Hartley's estimator by using additional information about N_{ab} , giving

$$\hat{Y}_{FB}(\beta) = \hat{Y}_a^A + \hat{Y}_b^B + \beta_1 \hat{Y}_{ab}^A + (1 - \beta_1) \hat{Y}_{ab}^B + \beta_2 (\hat{N}_{ab}^A - \hat{N}_{ab}^B).$$

As with Hartley's estimator, the optimal values β_{1opt} and β_{2opt} are chosen to minimize the variance of $\hat{Y}_{FB}(\beta)$, and are thus functions of the covariances of the domain totals. Substituting consistent estimators $\hat{\beta}_{1opt}$ and $\hat{\beta}_{2opt}$ gives the weight adjustments for w_i^A and w_i^B . Lohr and Rao (2000) showed that the Fuller-Burmeister estimator \hat{Y}_{FB} has the smallest asymptotic variance among the estimators considered. As with the Hartley estimator, however, the modified weights are random variables that differ for different responses, and in complex sampling designs the Fuller-Burmeister estimator is also internally inconsistent.

Pseudo-maximum likelihood (PML) estimators. To achieve internal consistency Skinner and Rao (1996) proposed a pseudo-maximum likelihood (PML) estimator that uses the same weights for all variables. When N_{ab} is unknown, it is estimated by $\hat{N}_{ab}^{PML}(\theta)$, which is the smaller of the roots of the quadratic equation

$$\left[\frac{\theta}{N_B} + \frac{1 - \theta}{N_A} \right] x^2 - \left[1 + \theta \frac{\hat{N}_{ab}^A}{N_B} + (1 - \theta) \frac{\hat{N}_{ab}^B}{N_A} \right] x + \theta \hat{N}_{ab}^A + (1 - \theta) \hat{N}_{ab}^B = 0.$$

Skinner and Rao (1996) suggested using the value θ_P for θ that minimizes the asymptotic variance of $\hat{N}_{ab}^{PML}(\theta)$:

$$\theta_P = \frac{N_a N_B V(\hat{N}_{ab}^B)}{N_a N_B V(\hat{N}_{ab}^B) + N_b N_A V(\hat{N}_{ab}^A)}. \quad (4)$$

Substituting an estimator $\hat{\theta}_P$ for θ_P , the weight adjustments are:

$$m_{i,P}^A = \begin{cases} \frac{N_A - \hat{N}_{ab}^{PML}(\hat{\theta}_P)}{\hat{N}_a^A} & \text{if } i \in a \\ \frac{\hat{N}_{ab}^{PML}(\hat{\theta}_P)}{\hat{\theta}_P \hat{N}_{ab}^A + (1 - \hat{\theta}_P) \hat{N}_{ab}^B} \hat{\theta}_P & \text{if } i \in ab, \\ \frac{N_B - \hat{N}_{ab}^{PML}(\hat{\theta}_P)}{\hat{N}_b^B} & \text{if } i \in b \\ \frac{\hat{N}_{ab}^{PML}(\hat{\theta}_P)}{\hat{\theta}_P \hat{N}_{ab}^A + (1 - \hat{\theta}_P) \hat{N}_{ab}^B} (1 - \hat{\theta}_P) & \text{if } i \in ab. \end{cases}$$

If the value of θ_P cannot be estimated, for example if the two sampling frames coincide or the design in Figure 1 is used, then one can use an average design effect from each survey in the adjustment, as described in Lohr and Rao (2006). The PML estimator is internally consistent; while

not guaranteed to give the smallest mean squared error, it has high efficiency in many survey situations.

Single frame estimators. Bankier (1986) and Kalton and Anderson (1986) proposed estimators of the form in (1) that treat all the observations as though they had been sampled from one frame, with adjusted weights in the intersection domain relying on the inclusion probabilities for each frame. The weight adjustments for the Kalton and Anderson (1986) single frame estimator are:

$$m_{i,S}^A = \begin{cases} 1 & \text{if } i \in a \\ w_i^B / (w_i^A + w_i^B) & \text{if } i \in ab, \end{cases}$$

$$m_{i,S}^B = \begin{cases} 1 & \text{if } i \in b \\ w_i^A / (w_i^A + w_i^B) & \text{if } i \in ab. \end{cases}$$

If $w_i^A = 1 / \pi_i^A$ and $w_i^B = 1 / \pi_i^B$, the single frame estimator uses $\tilde{w}_{i,S}^A = \tilde{w}_{i,S}^B = 1 / (\pi_i^A + \pi_i^B)$ for units in ab . The weight adjustment in domain ab relies on both π_i^A and π_i^B . Thus if a disproportionate stratified random sample is taken from frame B, one must know the frame-B stratum membership for units sampled in $S(A)$. The adjusted weights from the single frame estimator can be interpreted in terms of inclusion probabilities for sampled units. If the sampling fractions are small, $\tilde{w}_{i,S}^A$ is approximately $1/P$ (unit i is included in one of the samples). If each of $S(A)$ and $S(B)$ is self-weighting, then the single frame estimator reduces to (3).

The single frame weight modifications are the same for all response variables, so estimators are internally consistent. For complex surveys, however, single frame estimators may not be as efficient as the optimal or PML estimators. Their performance may be improved by raking toward the frame population totals (Skinner 1991).

Pseudo-empirical likelihood (PEL) estimators. Rao and Wu (2010) proposed empirical likelihood estimators for dual frame surveys. Using $\theta = \theta_P$, the empirical log likelihood function is defined by

$$\ell(\mathbf{p}_a, \mathbf{p}_{ab}^A, \mathbf{p}_{ab}^B, \mathbf{p}_b) = \frac{n_A + n_B}{N} \left[\sum_{i \in S(A), i \in a} \frac{N_a}{\hat{N}_a} w_i^A \log(p_{ai}) + \sum_{i \in S(A), i \in ab} \frac{\theta_P N_{ab}}{\hat{N}_{ab}^A} w_i^A \log(p_{abi}^A) + \sum_{i \in S(B), i \in b} \frac{N_b}{\hat{N}_b} w_i^B \log(p_{bi}) + \sum_{i \in S(B), i \in ab} \frac{(1 - \theta_P) N_{ab}}{\hat{N}_{ab}^B} w_i^B \log(p_{abi}^B) \right],$$

where θ_p is given in (4). An estimator $\hat{\theta}_p$ is substituted if θ_p is unknown. Then $\ell(\mathbf{p}_a, \mathbf{p}_{ab}^A, \mathbf{p}_{ab}^B, \mathbf{p}_b)$ is maximized subject to

$$\sum_{i \in S(A), i \in a} p_{ai} = 1, \sum_{i \in S(A), i \in ab} p_{abi}^A = 1, \\ \sum_{i \in S(B), i \in b} p_{bi} = 1, \sum_{i \in S(B), i \in ab} p_{abi}^B = 1,$$

and

$$\sum_{i \in S(A), i \in ab} p_{abi}^A y_i = \sum_{i \in S(B), i \in ab} p_{abi}^B y_i. \quad (5)$$

When N_{ab} is unknown, the PEL weight modifications are

$$m_{i, \text{PEL}}^A = \begin{cases} \frac{p_{ai}^A}{w_i^A} \{N_A - \hat{N}_{ab}^{\text{PML}}(\hat{\theta}_p)\} & \text{if } i \in a \\ \hat{\theta}_p \frac{p_{abi}^A}{w_i^A} \hat{N}_{ab}^{\text{PML}}(\hat{\theta}_p) & \text{if } i \in ab, \end{cases} \\ m_{i, \text{PEL}}^B = \begin{cases} \frac{p_{bi}^B}{w_i^B} \{N_B - \hat{N}_{ab}^{\text{PML}}(\hat{\theta}_p)\} & \text{if } i \in b \\ (1 - \hat{\theta}_p) \frac{p_{abi}^B}{w_i^B} \hat{N}_{ab}^{\text{PML}}(\hat{\theta}_p) & \text{if } i \in ab. \end{cases}$$

The constraint in (5) changes the weights in the overlap domain so that the estimator of Y_{ab} from $S(A)$ is forced to equal the estimator of Y_{ab} from $S(B)$. This constraint, however, results in a different set of weights for each response variable. The PEL estimator thus is not internally consistent. Rao and Wu (2010) presented an alternative multiplicity version in which the weight adjustments do not depend on y ; in the absence of auxiliary information, this estimator is the same as $\hat{Y}(1/2)$ in (3).

2.2 Weight adjustments with three or more frames

In the general case, suppose there are Q frames, denoted A_1, \dots, A_Q . Let $S(A_q)$ denote the probability sample from frame A_q , for $q = 1, \dots, Q$. Unit i in sample $S(A_q)$ has probability of inclusion $\pi_i^{A_q}$ and weight $w_i^{A_q}$. There are a total of D distinct domains.

A multiple frame estimator generalizing (1) is of the form

$$\hat{Y} = \sum_{q=1}^Q \sum_{i \in S(A_q)} m_i^{A_q} w_i^{A_q} y_i,$$

where $m_i^{A_q}$ is the weight adjustment for observation i in $S(A_q)$. A fixed weight estimator sets weight adjustments $m^{(A_q, d)}$ for each frame and domain, with the constraints that $m^{(A_q, d)} \geq 0$ ($m^{(A_q, d)}$ is assumed to equal 0 if domain d is not part of frame A_q) and $\sum_{q=1}^Q m^{(A_q, d)} = 1$ for $d = 1, \dots, D$. Then, $m_i^{A_q} = m^{(A_q, d)}$ when observation i from $S(A_q)$ is in domain d . A simple choice, which generalizes the fixed weight dual frame estimator $\hat{Y}(1/2)$ in (3), takes $m^{(A_q, d)} = [1/\text{number of frames that contain domain } d]$; this is called the multiplicity estimator by Mecatti (2007). Other choices include setting

$m^{(A_q, d)} = 1$ in exactly one frame and 0 for the other frames, resulting in screening estimators.

Many of the properties from the dual frame situation extend to the case of three or more frames; multiple frame versions of the estimators in Section 2.1 were studied by Hartley (1974), Lohr and Rao (2006), and Mecatti (2007). How do the multiple frame estimators satisfy the criteria set out at the beginning of this section? All of the estimators – fixed weight, optimal, PML, PEL, and single frame – are approximately unbiased for population totals when sufficiently large samples are taken in the frames. The fixed weight, PML, and single frame estimators are internally consistent; the optimal Hartley-type and Fuller-Burmeister-type estimators in Lohr and Rao (2006) and a multiple-frame extension of the PEL estimator of Rao and Wu (2010) are not internally consistent. While the optimal estimators are asymptotically efficient, they are often unstable in small or moderate samples with three or more frames because the optimal estimated weight modifications are functions of large estimated covariance matrices. The optimal and PEL estimators are ill suited for use with standard survey software because they require a different set of weights for each response variable.

We recommend that one of the internally consistent estimators – fixed weight, PML, or single frame – be used in practice. Lohr and Rao (2006) concluded that the PML estimator has small mean squared error in many survey circumstances, and thus is a good choice for a survey that is conducted only once. With repeated surveys, though, the simplicity and transparency of a fixed weight estimator may be preferred. Fixed weight adjustments may make year-to-year comparisons easier in an annual survey where the domain proportions are relatively constant over time. Fixed weight estimators are also more amenable to weight adjustments for nonresponse and domain misclassification (see Sections 5.1 and 6.1). If fixed weight adjustments can be chosen that are close to the optimal weight adjustments for important responses, perhaps by using estimated design effects from previous surveys, the fixed weight estimator will have mean squared error close to that of the optimal and PML estimators.

3. Postratification to population controls

All of the estimators in Section 2 modify the original sampling weights. As a result, some properties of the original weights may be lost. For example, if a stratified random sample is taken in frame A, the modified weights will not necessarily have the property that the sum of the weights in a stratum equals the stratum population size.

Bankier (1986), in the original development of single frame estimation methods, suggested raking the single

frame weights, $\tilde{w}_{i,S}^A$ and $\tilde{w}_{i,S}^B$, to stratum totals so that the adjusted weights $\tilde{w}_{i,S}^{A,adj}$ and $\tilde{w}_{i,S}^{B,adj}$ satisfy

$$\sum_{i \in S_{Ah}} (\tilde{w}_{i,S}^{A,adj} + \tilde{w}_{i,S}^{B,adj}) = N_{Ah},$$

where S_{Ah} represents the sampled units from either frame in stratum h of frame A, and N_{Ah} is the population size of that stratum. Bankier (1986) and Skinner (1991) used raking ratio estimation to calibrate single frame estimators to the frame population sizes N_A and N_B . Kott, Amrhein and Hicks (1998) proposed using the least squares calibration methods of Deville and Särndal (1992) for calibrating weights to population totals such as stratum sizes.

For the PML estimator, Lohr and Rao (2000) recommended combining the samples first and then using calibration methods to adjust to population as well as separate-frame population totals. When nonresponse is present and a fixed weight estimator is used, Brick, Cervantes, Lee and Norman (2011) concluded that it is preferable to poststratify the individual samples first, and then combine the samples. In some situations, it is most efficient to poststratify both before and after combining samples; in other situations, poststratification can increase bias (see Section 6). Decisions about poststratification need to be made based on the mean squared error, which includes effects of nonsampling errors, and not just on the sampling variance.

4. Analyzing multiple frame surveys with survey software

4.1 Point estimation with survey software

Only internally consistent weight adjustments are suitable for use with survey software when there are multiple responses of interest. Each of the internally consistent methods presented in Section 2.1 results in one vector of adjusted weights for each sample. These may then be concatenated to form one vector of weights: $\tilde{\mathbf{w}} = [\tilde{w}_i^A, i \in S(A_1), \dots, \tilde{w}_i^B, i \in S(A_Q)]$. Let \mathbf{y} be the corresponding vector of observations, formed by concatenating the observations from samples $S(A_1)$ through $S(A_Q)$. Then $\hat{Y} = \tilde{\mathbf{w}}' \mathbf{y}$. From a user's perspective, once the modified weights are constructed, the procedure followed to find point estimates of population totals and means is the same as in a single frame survey.

The modified weights from an internally consistent procedure can be used to estimate any population quantity. Let $F(y)$ be the cumulative distribution function for the population, with

$$F(y) = \sum_{i=1}^N I(y_i \leq y) / N,$$

where $I(y_i \leq y) = 1$ if $y_i \leq y$ and 0 otherwise. In a single frame survey, $F(y)$ is estimated by the empirical cumulative distribution function

$$\hat{F}(y) = \sum_{i \in S} w_i I(y_i \leq y) / \sum_{i \in S} w_i.$$

The modified weights may be used to estimate $F(y)$ in a multiple frame survey:

$$\hat{F}(y) = \frac{\sum_{q=1}^Q \sum_{i \in S(A_q)} \tilde{w}_i^{Aq} I(y_i \leq y)}{\sum_{q=1}^Q \sum_{i \in S(A_q)} \tilde{w}_i^{Aq}}.$$

The denominator is approximately unbiased for N , and the numerator is approximately unbiased for $\sum_{i=1}^N I(y_i \leq y)$. Any functional of the cumulative distribution function may then be estimated using $\hat{F}(y)$: the mean, $\int y dF(y)$, the median m satisfying $F(m) \approx 1/2$, or any other quantity.

Since the estimators with modified weights are approximately unbiased for population means and totals, they are also approximately unbiased for smooth functions of population means such as ratios and regression coefficients. Any population quantity that could be estimated using the weights from a single frame survey can be estimated analogously using the adjusted weight vector for the multiple frame survey.

4.2 Variance estimation with survey software

Knowledge of the survey designs is needed to calculate standard errors. Variance estimation is straightforward for the estimator in (3), where the weight adjustments do not depend on the data. In that situation,

$$V[\hat{Y}(\theta)] = V\left[\sum_{i \in S(A)} \tilde{w}_i^A y_i\right] + V\left[\sum_{i \in S(B)} \tilde{w}_i^B y_i\right],$$

where \tilde{w}_i^A and \tilde{w}_i^B are defined below (2). Create the data set by concatenating the observations from $S(A)$ and $S(B)$ as in Section 4.1, using \tilde{w}_i^A and \tilde{w}_i^B as the weights. Define the stratification variable for the combined sample as the combination of categories given by the frame indicator variable, the frame-A stratification variable, and the frame-B stratification variable. Define the first-stage clustering variable for the combined sample similarly as the combination of categories of the individual frame clustering variables. Then, standard survey software may be used to estimate population means and totals using the modified weights, and to estimate variances using the stratification and clustering variables from the combined samples.

Variance estimation is more complicated when the weight modifications m_i^A or m_i^B depend on quantities that are estimated from the sample, as in the PML estimator, or when the combined sample is poststratified or calibrated to population quantities. Linearization, jackknife, and bootstrap methods may then be used to estimate variances.

In the following, we summarize methods that can be used for variance estimation if the psus from the frames are selected independently. If samples from the different frames share psus, other methods must be used. If, for example, psus are selected from the population, and a dual frame design is used within each selected psu, point estimators for psu totals can be calculated using one of the methods described in Section 2. Then standard replication methods can be used to calculate a with-replacement variance estimator.

Under regularity conditions, the linearization and jackknife methods are consistent for estimating the variance of a population characteristic τ that can be written as $\tau = g(\mathbf{A}, \mathbf{B})$, where \mathbf{A} is a vector of population totals from frame A, \mathbf{B} is a vector of population totals from frame B, and g is a twice continuously differentiable function (Skinner and Rao 1996; Lohr and Rao 2000). The vector \mathbf{A} is estimated from $S(A)$ by $\hat{\mathbf{A}}$, with estimated covariance matrix $\hat{\Sigma}_A$; similarly, $\hat{\mathbf{B}}$ estimates \mathbf{B} from $S(B)$, with $\hat{V}(\hat{\mathbf{B}}) = \hat{\Sigma}_B$. The linearization estimator of the variance of $\hat{\tau} = g(\hat{\mathbf{A}}, \hat{\mathbf{B}})$ is

$$\hat{V}_L(\hat{\tau}) = g'_A \hat{\Sigma}_A g_A + g'_B \hat{\Sigma}_B g_B,$$

where g_A is the vector of partial derivatives of $g(\mathbf{A}, \mathbf{B})$ with respect to the components of \mathbf{A} and g_B is the corresponding vector of partial derivatives for frame B. Demnati, Rao, Hidiroglou and Tambay (2007) derived linearization estimators of the variance for multiple frame surveys by taking derivatives of a function of the weights rather than of the means. Linearization methods require that the derivatives be calculated separately for each estimator that is considered, and these calculations can be cumbersome. For that reason, it may be preferred to use replication methods if multiple frame surveys are adopted.

Suppose a stratified multistage sample with H strata is taken from frame A, where stratum h has \tilde{n}_h^A primary sampling units. An independent stratified multistage sample with L strata is taken from frame B, where stratum l has \tilde{n}_l^B primary sampling units. The jackknife estimator of the variance can be calculated by creating a total of $\sum_{h=1}^H \tilde{n}_h^A + \sum_{l=1}^L \tilde{n}_l^B$ replicate weight columns (Lohr and Rao 2000). The replicate weights for the column corresponding to the deletion of psu i from stratum h in S_A are formed by:

$$\tilde{w}_{k(hi)}^A = \begin{cases} \frac{\tilde{n}_h^A}{\tilde{n}_h^A - 1} \tilde{w}_k^A & \text{if unit } k \text{ is in stratum } h \text{ but not in psu } i, \\ 0 & \text{if unit } k \text{ is in psu } i \text{ of stratum } h, \\ \tilde{w}_k^A & \text{if unit } k \text{ is in stratum } g \neq h. \end{cases}$$

The jackknife coefficient for this column is the multiplier $(\tilde{n}_h^A - 1) / \tilde{n}_h^A$. The column of replicate weights corresponding to the deletion of psu j from stratum l in S_B is

formed similarly, with jackknife coefficient $(\tilde{n}_l^B - 1) / \tilde{n}_l^B$. With more than two frames, additional columns of replicate weights are added corresponding to the deleted psus from those samples. Weights for a bootstrap method of variance estimation (see Lohr 2007) can be defined similarly.

Multiple frame replication variance methods can be used with standard survey packages that allow replicate weights. If desired, each column in the replicate weights can be post-stratified to population and frame totals, so that the post-stratification is accounted for in the variance estimation.

One challenge with replication variance methods is that the number of columns of replicate weights needed may be very large if a simple random sample or stratified random sample is taken in one of the frames. For the bootstrap, we have found that for some surveys at least 500 bootstrap iterations are needed for variance estimates with dual frame surveys, which again may be excessive. It is possible that combined strata variance estimation, as discussed in Lu, Brick and Sitter (2006), may be used with multiple frame surveys to reduce the number of replicates needed.

5. Nonsampling errors

Multiple frame surveys often have better population coverage than a single frame surveys. When all frames are incomplete, as in Figure 3, any one of frames A, B, or C, if used as the sole sampling frame, would have severe undercoverage. The multiple frame survey design ensures that all units in the overlapping frames have a positive probability of inclusion.

Like all surveys, multiple frame surveys are subject to nonsampling errors. They have nonresponse, which may differ in the different frames. While the union of the frames may have better coverage than a single frame, there may still be undercoverage of the target population. Estimators for multiple frame surveys are also sensitive to domain misclassification and biases that might result from different administration methods or modes in the component surveys. We discuss nonresponse and mode effects in this section, and study effects of domain misclassification in Section 6.

5.1 Nonresponse

In any survey, nonresponse can result in biased estimates of population totals and other quantities. Different nonresponse rates in the samples from the two frames can affect the point estimates of the population total given in Section 2; additionally, nonresponse can affect the weight adjustments prescribed by some of the methods.

Kennedy (2007) discussed a problem that has occurred when frame A consists of landline telephone numbers and frame B has cellular telephone numbers: the units in the

intersection domain ab who were interviewed by cell phone differed from those in ab who were interviewed on the landline phone. For example, it was estimated that 18% of the intersection units were aged 18-25 in the frame-B sample, while it was estimated that only 8% of the intersection units were aged 18-25 using the frame-A sample. The difference was ascribed to nonresponse: it was thought that persons who predominantly use cellular telephones (and thus are difficult to reach through a landline survey) tend to be younger. Kennedy (2007) suggested raking using estimated relative telephone usage (*i.e.*, whether most of calls are on landline or cellular telephone).

Brick *et al.* (2011) proposed two methods for non-response adjustment in dual frame cellular/landline telephone surveys with fixed weight estimators. They considered a setup in which the overlap domain has two groups: households that receive all or nearly all of their calls on cellular telephones (cell-mainly), and the remaining households in the overlap domain (landline-mainly). The first method, which does not require external estimates of control totals, sets the value of θ in the fixed weight adjustment estimator to reduce the nonresponse bias by using the response rates for the cell-mainly and landline-mainly households in each sample. The second method requires poststratification control totals for the cell-mainly and landline-mainly groups in the overlap domain, N_{1ab} and N_{2ab} , and estimates the population total in domain ab by

$$\sum_{g=1}^2 \left[\theta_g \frac{N_{gab}}{\hat{N}_{gab}^A} \hat{Y}_{gab}^A + (1 - \theta_g) \frac{N_{gab}}{\hat{N}_{gab}^B} \hat{Y}_{gab}^B \right],$$

where \hat{Y}_{gab}^A represents the estimated total of group g in domain ab from $S(A)$, the other totals are defined similarly, and $0 \leq \theta_g \leq 1$ for $g = 1, 2$.

5.2 Mode effects

In some cases, multiple frame may also mean multiple mode. De Leeuw (2008) compared the advantages and disadvantages of different sampling modes, and summarized empirical research on mode biases. Persons may give different responses when presented with questions in a visual form than when presented with questions in an auditory form, resulting in mode bias. Mode effects that occur in single frame surveys will also occur in multiple frame surveys. If different modes are used in different frames, it is challenging to separate mode effects from other nonsampling errors.

Many of the multiple frame survey estimators combine estimates from the overlap domains, and these methods assume that the estimators of Y_{ab} from the component surveys both estimate the same quantity. If, however, the frame A survey is conducted in person while the frame B

survey is conducted by telephone, it is possible that a census of the domain ab from frame B would give a different domain total than a census from frame A.

One possibility to investigate mode effects is to conduct the frame B survey using a split sample, *e.g.*, partly in person and partly by telephone, but that would reduce the cost savings from the dual frames. Careful pretesting can mitigate the mode effects. Research is needed in this area; the same problem of mode effects, of course, occurs in single frame surveys such as the American Community Survey in which nonresponse follow-up is done by different mode than the original sample (see Citro and Kalton 2007). The methods presented in de Leeuw, Hox and Dillman (2008) for designing surveys for multiple modes also apply in the multiple frame setting.

Vannieuwenhuyze, Loosveldt and Molenberghs (2011) presented a method for distinguishing mode effects from selection effects when a supplemental single-mode survey is available. They noted, however, that the method requires the strong assumption that the coverage and nonresponse errors are equivalent for both surveys. If this assumption is met for a dual frame survey so that the samples in the overlap domain from frames A and B represent the same population, and if domain classification is correct, the mode effect can be estimated from the overlap domain as $D_{ab} = \hat{Y}_{ab}^A - \hat{Y}_{ab}^B$. A difference that is significantly different from 0 indicates presence of a mode effect if there are no other nonsampling errors. If other nonsampling errors are present, a large value of D_{ab} does not provide information about the cause of the difference; experimentation is needed to distinguish possible causes.

6. Domain misclassification and bias adjustment

The estimators discussed in Section 2 construct weights for the observations based on domain membership. Thus in the estimator $\hat{Y}(\theta)$ in (3), the weight multiplier of an observation from sampling frame A is 1 if the observation is in domain a , and is θ if the observation is in domain ab , in order to account for the multiplicity of sampling.

In practice, domain membership may not be clear. For the situation in Figure 1, it may be unknown whether a respondent in an area frame also belongs to the list frame. If frame A is an area frame and frame B is an internet frame, for example, the only way to determine whether an individual sampled from frame A is also in frame B may be to ask the person about internet access, and the person might not give the correct response.

If matching or record linkage is used to determine frame membership, imperfect matching can also misclassify observations. Lesser and Kalsbeek (1999) discussed nonsampling errors that occur in dual frame surveys that have been

conducted by the U.S. National Agricultural Statistics Service. Domain misclassification can occur if a farm sampled in the area frame is incorrectly classified with respect to its list frame membership. In landline/cellular dual frame telephone surveys, it is challenging to determine whether a person in one frame is also in the other frame (Kennedy 2007). A person reached in a landline telephone sample may also have a cell phone, but rarely take calls on the cell phone. While technically in the overlap domain, that person is virtually unreachable in the cell phone survey. Some landline/cellular surveys ask respondents about the relative amounts of cellular or landline telephone usage, but misclassification can occur.

In practice, we expect domain misclassification to be related to responses of interest; we also expect that in many situations, misclassification is more likely to occur in certain directions. In longitudinal dual frame surveys, domain misclassification can have greater effects than in cross-sectional surveys (Lu and Lohr 2010). In some situations, the domain indicator can be missing or unavailable. Clark, Winglee and Liu (2007) investigated logistic regression and record-linkage methods for predicting the domain of an observation with missing domain information.

6.1 Misclassification bias adjustments

If domain misclassification is severe, each method for modifying the survey weights to adjust for multiplicity can result in biased estimates of population quantities. In this section we derive a correction for the domain misclassification bias of the fixed weight estimator of Section 2.2 when misclassification probabilities are known. Let the D -vector $\delta_i^{A_q}$ denote the true domain membership for observation i of frame A_q , containing a 1 in position d if observation i is in domain d , and 0 elsewhere. Let $\mathbf{Y} = (Y_1, \dots, Y_D)'$ denote the vector of population totals for the D domains. For an overlapping dual frame survey, $\mathbf{Y} = (Y_a, Y_{ab}, Y_b)'$; for a three-frame survey, $\mathbf{Y} = (Y_a, Y_{ab}, Y_{ac}, Y_{abc}, Y_b, Y_{bc}, Y_c)'$. If there is no domain misclassification,

$$\hat{\mathbf{Y}}^{A_q} = \sum_{i \in S(A_q)} \delta_i^{A_q} w_i^{A_q} y_i$$

is the corresponding estimator of \mathbf{Y} from $S(A_q)$. For fixed weight adjustment vector $\mathbf{m}^{A_q} = (m^{(A_q, 1)}, \dots, m^{(A_q, D)})'$ in frame A_q , satisfying $\sum_{q=1}^Q \mathbf{m}^{(A_q, d)} = 1$, then $E[\sum_{q=1}^Q (\mathbf{m}^{A_q})' \hat{\mathbf{Y}}^{A_q}] = \mathbf{Y}$.

Now suppose there is misclassification. Let $\eta_i^{A_q}$ denote the observed classification for observation i in S . We can write $\eta_i^{A_q} = (\mathbf{M}_i^{A_q})' \delta_i^{A_q}$, where $\mathbf{M}_i^{A_q}$ is a $D \times D$ matrix containing a 1 in position (d, e) if observation i in true domain d is (mis)classified to domain e , and 0 elsewhere.

To allow for differential misclassification within domains, we posit a structure in which the misclassification probabilities can differ for subpopulations in a frame. In a landline/cellular survey, for example, some age groups may be known to have higher misclassification probabilities than others. Chambers, Chipperfield, Davis and Kovačević (2008) used a similar grouping approach to correct for record linkage errors. Suppose the population can be divided into G groups, $g = 1, \dots, G$, in which the misclassification probabilities are known for each frame A_q . Let $\phi_g^{A_q}(d, e)$ denote the probability that an observation in group g with true domain d is classified into domain e in sample $S(A_q)$, and let $\Phi_g^{A_q}$ be the $D \times D$ matrix with entries $\phi_g^{A_q}(d, e)$. For observation i belonging to group g and true domain d , assume that row d of $\mathbf{M}_i^{A_q}$ is generated as a multinomial random variable of size 1 with probabilities in row d of the expected misclassification matrix $\Phi_g^{A_q}$, and that all $\mathbf{M}_i^{A_q}$ are independent of each other and of the sample inclusion variables. We thus have G matrices of misclassification probabilities for frame A_q , $\Phi_1^{A_q}, \dots, \Phi_G^{A_q}$. Denote the vector of population totals for group g by $\mathbf{Y}(g) = \sum_{i=1}^N \delta_i^{A_q} \chi_i(g) y_i$, where $\chi_i(g) = 1$ if observation i is in group g and 0 otherwise.

With the observed domain classifications $\eta_i^{A_q}$, the design-weighted estimator of the vector of domain totals in group g is

$$\begin{aligned} \hat{\mathbf{Y}}^{A_q}(\text{mis}, g) &= \sum_{i \in S(A_q)} \eta_i^{A_q} \chi_i(g) w_i^{A_q} y_i \\ &= \sum_{i \in S(A_q)} (\mathbf{M}_i^{A_q})' \delta_i^{A_q} \chi_i(g) w_i^{A_q} y_i, \end{aligned}$$

so that $E[\hat{\mathbf{Y}}^{A_q}(\text{mis}, g)] = (\Phi_g^{A_q})' \mathbf{Y}(g)$.

Now consider a new vector of weight adjustments $\tilde{\mathbf{m}}_g^{A_q} = (\tilde{m}_g^{(A_q, 1)}, \dots, \tilde{m}_g^{(A_q, D)})'$ for group g in frame A_q . Then

$$E \left[\sum_{g=1}^G \sum_{q=1}^Q (\tilde{\mathbf{m}}_g^{A_q})' \hat{\mathbf{Y}}^{A_q}(\text{mis}, g) \right] = \sum_{g=1}^G \sum_{q=1}^Q (\Phi_g^{A_q} \tilde{\mathbf{m}}_g^{A_q})' \mathbf{Y}(g).$$

Since $\sum_{g=1}^G \sum_{q=1}^Q (\mathbf{m}^{A_q})' \mathbf{Y}(g) = \mathbf{Y}$, the bias will be eliminated under this model when

$$\tilde{\mathbf{m}}_g^{A_q} = (\Phi_g^{A_q})^+ \mathbf{m}^{A_q}, \quad (6)$$

where $(\Phi_g^{A_q})^+$ is the Moore-Penrose inverse of $\Phi_g^{A_q}$, obtained by taking the inverse of the nonzero rows and columns of $\Phi_g^{A_q}$.

Replacing weight adjustments \mathbf{m}^{A_q} by $\tilde{\mathbf{m}}_g^{A_q}$ eliminates the bias under the multinomial misclassification model but inflates the variance. For frame A_q ,

$$\begin{aligned}
& V \left[\sum_{g=1}^G (\tilde{\mathbf{m}}_g^{A_q})' \hat{\mathbf{Y}}^{A_q}(\text{mis}, g) \right] \\
&= E \left[V \left(\sum_{g=1}^G \sum_{i \in S(A_q)} \{(\Phi_g^{A_q})^+ \mathbf{m}_i^{A_q}\}' (\mathbf{M}_i^{A_q})' \right. \right. \\
&\quad \left. \left. \delta_i^{A_q} \chi_i(g) w_i^{A_q} y_i \mid S(A_1), \dots, S(A_Q) \right) \right] \\
&+ V \left[E \left(\sum_{g=1}^G \sum_{i \in S(A_q)} \{(\Phi_g^{A_q})^+ \mathbf{m}_i^{A_q}\}' (\mathbf{M}_i^{A_q})' \right. \right. \\
&\quad \left. \left. \delta_i^{A_q} \chi_i(g) w_i^{A_q} y_i \mid S(A_1), \dots, S(A_Q) \right) \right] \\
&= \sum_{g=1}^G [(\Phi_g^{A_q})^+ \mathbf{m}_i^{A_q}]' E \left[\sum_{i \in S(A_q)} \chi_i(g) (w_i^{A_q} y_i)^2 \right. \\
&\quad \left. \{ \text{diag}[(\Phi_g^{A_q})' \delta_i^{A_q}] - (\Phi_g^{A_q})' \delta_i^{A_q} (\delta_i^{A_q})' \Phi_g^{A_q} \} \right] (\Phi_g^{A_q})^+ \mathbf{m}_i^{A_q} \\
&+ V \left[\sum_{i \in S(A_q)} \{ \mathbf{m}_i^{A_q} \}' \delta_i^{A_q} w_i^{A_q} y_i \right].
\end{aligned}$$

The second term is the variance of the contribution from frame A_q when the units are classified correctly. The first term is zero only when $\Phi_g^{A_q}$ is diagonal for all g , i.e., there is no misclassification.

The weight adjustments in (6) may be extended to the case in which the original fixed weights $\mathbf{m}_i^{A_q}$ vary for the groups, as long as $\sum_{q=1}^Q m_g^{(A_q, d)} = 1$ for each domain. Note that the bias correction method in this section is proposed only for the fixed weight estimators, and not for the PML, PEL, or optimal estimators where the multiplicity weights depend on the data. The bias correction depends on the correct specification of the misclassification probabilities. If the misclassification probabilities are estimated from another survey, the operational methods of the surveys must be similar.

6.2 Simulation study

Lohr and Rao (2006) found in simulation studies that the PML estimator has smaller mean squared error than the other estimators when random misclassification is present, but this is due largely to the smaller variance of that estimator. To study sensitivity of estimators to other forms of domain misclassification, we performed a simulation study for two- and three-frame surveys. The population for domain d was generated using the model $y_{ij} = \mu_d + \alpha_i + \varepsilon_{ij}$ for $i = 1, \dots, N_d$ and $j = 1, \dots, 5$, with $\alpha_i \sim N(0, 1)$ and $\varepsilon_{ij} \sim N(0, 1)$ generated independently, and then probability samples were drawn from this population.

For the two-frame study, the domain means are $\mu_a = -1$, $\mu_{ab} = 0$, $\mu_b = 2$ and factors in the simulation are:

1. Sample size: 100 or 200 from each frame.
2. Cluster sample or simple random sample drawn from frame A. A cluster sample was drawn by

selecting a simple random sample of $n_A/5$ of the groups used in generating the population.

3. Misclassification probabilities for frame A (all probabilities not listed are 0):
 - a. $\phi_{aa}^A = 1$, $\phi_{ab,ab}^A = 1$ (no misclassification);
 - b. $\phi_{aa}^A = 0.9$, $\phi_{a,ab}^A = 0.1$, $\phi_{ab,ab}^A = 1$;
 - c. $\phi_{aa}^A = 0.9$, $\phi_{a,ab}^A = 0.1$, $\phi_{ab,ab}^A = 0.9$, $\phi_{ab,a}^A = 0.1$;
 - d. $\phi_{aa}^A = 1$, $\phi_{ab,ab}^A = 0.9$, $\phi_{ab,a}^A = 0.1$.
4. Misclassification probabilities for frame B:
 - a. $\phi_{bb}^B = 1$, $\phi_{b,ab}^B = 1$ (no misclassification);
 - b. $\phi_{bb}^B = 0.8$, $\phi_{b,ab}^B = 0.2$, $\phi_{ab,ab}^B = 1$;
 - c. $\phi_{bb}^B = 0.8$, $\phi_{b,ab}^B = 0.2$, $\phi_{ab,ab}^B = 0.9$, $\phi_{ab,b}^B = 0.1$;
 - d. $\phi_{bb}^B = 1$, $\phi_{ab,ab}^B = 0.8$, $\phi_{ab,b}^B = 0.2$.
5. Population sizes: $N_a = N_b = N_{ab} = 25,000$; $N_a = N_b = 10,000$, $N_{ab} = 55,000$; $N_a = 25,000$, $N_{ab} = 40,000$, $N_b = 10,000$.

Ten thousand replicates were run for each combination of the factors, giving the Monte Carlo estimate of bias a standard error of approximately 100. We studied all estimators in Section 2, including $\hat{Y}(1/2)$, $\hat{Y}(2/3)$, and $\hat{Y}(1)$ from (3). We also examined poststratified estimators that could be employed when the domain population counts N_a , N_{ab} , and N_b are known: estimators with subscript “post1” apply poststratification to the two samples first and then combine the samples, and estimators with subscript “post2” combine the samples first and then poststratify to the domain population counts. The bias corrected estimators $\hat{Y}(1/2)_{bc}$ and $\hat{Y}(2/3)_{bc}$ modify the initial fixed weights corresponding to $\theta = 1/2$ and $\theta = 2/3$ using (6). With misclassification pattern (b) in frame A, for example, the bias-corrected weight adjustments for $\hat{Y}(1/2)_{bc}$ are $\tilde{m}_i^A = 19/18$ for i classified in a and $\tilde{m}_i^A = 1/2$ for i classified in ab ; for pattern (c), the bias-corrected weight adjustments are $17/16$ and $7/16$, respectively. The single frame estimator is omitted from these tables since it is the same as either $\hat{Y}(1/2)$ or $\hat{Y}(2/3)$; the single frame estimator raked to the population totals N_A and N_B is denoted by $\hat{Y}_{SF, \text{rake}}$. Tables 1 and 2 display results for $n_A = 100$, $n_B = 100$, $N_a = N_{ab} = N_b = 25,000$, and a simple random sample from frame A; Tables 3 and 4 give results for $n_A = 200$, $n_B = 100$, $N_a = N_{ab} = N_b = 25,000$, and a cluster sample from frame A. The general patterns of results are similar for the other simulations and are not shown here.

First, consider the fixed weight estimators. The bias-corrected estimators reduce the bias as expected; in all cases studied with misclassification, the empirical bias from the bias-corrected estimators was less than 200 in absolute value, which is within the margin of error. Although the standard deviation for the bias-corrected estimators is higher

than for the uncorrected estimators, in most cases the mean squared errors are comparable.

The screening estimator $\hat{Y}(1)$, which discards units from frame B in domain ab , exhibits no misclassification bias when frame B is correctly classified. It also exhibits no bias in Tables 1 and 3 with frame-B misclassification pattern (d) because the observations misclassified from domain ab to domain b have mean 0; for different sets of domain means, pattern (d) does create bias. For the other cases, the screening estimator has the highest bias. For every misclassification pattern, the screening estimator has high mean squared error because data are thrown away. If the domain means are similar, then the misclassification might not result in appreciable bias but discarding observations from domain ab in $S(B)$ would greatly increase the mean squared error.

Poststratifying to the domain totals when there is misclassification often increases the bias instead of decreasing it. Consider line 4 of Table 1, where 20% of the $S(B)$ observations in ab are mistakenly classified into domain b . The weights of the observations that are really in domain b , with mean 2, are reduced from 500 to approximately 417, which causes the poststratified versions of $\hat{Y}(1/2)$ to be biased. The effect of poststratification on the mean squared error is mixed, and depends on whether the variance

reduction achieved by poststratifying exceeds the additional bias that can be introduced. Raking to the frame totals N_A and N_B , in $\hat{Y}_{SF, rake}$, has similar effect on misclassification bias as poststratification.

For the simple random samples in Tables 1 and 2, the PML and PEL estimators often exhibit much more bias than the uncorrected fixed weight estimators. The relative contributions from the two frames for these methods depend on the estimated variances of \hat{N}_{ab}^A and \hat{N}_{ab}^B , the domain weights depend on \hat{N}_{ab}^{PML} , and these two factors interact in complex ways depending on the misclassification structure. For misclassification pattern (d) in either frame, \hat{N}_{ab}^{PML} is too small because observations in domain ab are misclassified; consequently, the weights for the observations in the nonoverlapping domains are too large. A poststratified version of the PML estimator shared the bias problems of the fixed weight poststratified estimators. The PEL estimator, by forcing the estimators of Y_{ab} to be equal, can worsen the bias. For example, in the simulation in line 3 of Table 1, with correct classification for frame A and pattern (c) for frame B, the PEL bias is 50% larger than the PML bias. In this case, the PEL estimator pulls the unbiased estimator \hat{Y}_{ab}^A from $S(A)$ toward the biased estimator from frame B. The optimal estimators also exhibit high bias.

Table 1
Estimated bias for dual frame misclassification, with $n_A = n_B = 100$ and a simple random sample taken from each frame. MPA and MPB refer to the misclassification patterns for frames A and B

MPA	MPB	$\hat{Y}(1/2)$	$\hat{Y}(1/2)_{post1}$	$\hat{Y}(1/2)_{post2}$	$\hat{Y}(1/2)_{bc}$	$\hat{Y}(2/3)$	$\hat{Y}(2/3)_{bc}$	$\hat{Y}(1)$	\hat{Y}_H	\hat{Y}_{FB}	\hat{Y}_{PML}	\hat{Y}_{PEL}	$\hat{Y}_{SF, rake}$
a	a	-194	-87	-87	-194	-215	-215	-258	-68	10	-121	-119	-163
a	b	-5,015	4,145	4,529	5	-6,678	17	-10,002	-5,417	1,248	2,486	1,542	2,361
a	c	-5,142	-1,118	-898	-128	-6,823	-138	-10,185	-5,413	-2,583	-1,650	-2,482	-1,690
a	d	-57	-8,430	-8,431	-47	-69	-55	-92	30	-6,576	-6,723	-6,725	-6,795
b	a	1,163	-1,238	-1,290	-82	748	-82	-82	1,355	-2,376	-2,631	-2,551	-2,704
b	b	-3,724	3,040	3,264	43	-5,784	65	-9,905	-3,967	-920	-30	-850	-100
b	c	-3,882	-2,192	-2,187	-124	-5,977	-136	-10,167	-3,954	-4,319	-3,821	-4,477	-3,853
b	d	1,322	-9,445	-9,621	92	917	104	108	1,600	-8,219	-8,720	-8,531	-8,879
c	a	1,366	1,315	1,312	123	969	140	174	1,530	1,529	1,325	1,355	1,276
c	b	-3,729	5,456	5,948	51	-5,801	64	-9,945	-4,216	2,096	3,500	2,391	3,355
c	c	-3,797	235	512	-15	-5,868	2	-10,011	-4,089	-1,377	-417	-1,318	-466
c	d	1,285	-7,072	-7,212	56	873	60	48	1,535	-4,665	-5,131	-4,976	-5,222
d	a	-120	2,134	2,134	-111	-132	-126	-155	32	3,710	3,535	3,538	3,470
d	b	-4,979	6,497	7,086	34	-6,620	65	-9,901	-5,599	4,339	5,928	4,788	5,697
d	c	-5,152	1,174	1,644	-137	-6,835	-152	-10,200	-5,622	310	1,626	578	1,540
d	d	90	-5,999	-5,998	107	98	119	114	193	-2,964	-3,116	-3,120	-3,155

Table 2

Estimated $\sqrt{\text{MSE}}$ for dual frame misclassification, with $n_A = n_B = 100$ and a simple random sample taken from each frame. MPA and MPB refer to the misclassification patterns for frames A and B

MPA	MPB	$\hat{Y}(1/2)$	$\hat{Y}(1/2)_{\text{post1}}$	$\hat{Y}(1/2)_{\text{post2}}$	$\hat{Y}(1/2)_{bc}$	$\hat{Y}(2/3)$	$\hat{Y}(2/3)_{bc}$	$\hat{Y}(1)$	\hat{Y}_H	\hat{Y}_{FB}	\hat{Y}_{PML}	\hat{Y}_{PEL}	$\hat{Y}_{SF, \text{rake}}$
a	a	9,646	7,917	7,910	9,646	9,729	9,729	10,304	9,677	8,151	8,081	8,115	8,075
a	b	10,602	9,351	9,531	9,926	11,531	10,197	14,181	11,157	8,212	8,377	8,198	8,311
a	c	10,779	8,622	8,603	10,071	11,715	10,402	14,376	11,243	8,817	8,514	8,720	8,508
a	d	9,789	11,719	11,704	9,674	9,884	9,795	10,432	9,819	10,979	10,978	11,003	11,007
b	a	9,623	8,182	8,185	9,718	9,686	9,766	10,307	9,780	8,446	8,447	8,444	8,459
b	b	9,955	9,054	9,137	9,995	10,949	10,212	14,069	10,489	8,074	7,913	7,995	7,898
b	c	10,146	9,014	9,014	10,160	11,197	10,489	14,404	10,616	9,443	9,108	9,448	9,114
b	d	9,868	12,600	12,716	9,826	9,952	9,927	10,567	10,023	12,063	12,284	12,188	12,371
c	a	9,843	8,185	8,180	9,887	9,853	9,877	10,341	9,991	8,516	8,417	8,442	8,402
c	b	10,049	10,113	10,396	10,039	11,029	10,229	14,127	10,662	8,520	8,863	8,529	8,778
c	c	10,247	8,701	8,718	10,254	11,233	10,534	14,306	10,799	8,762	8,527	8,669	8,516
c	d	10,021	10,861	10,936	9,966	10,068	10,016	10,579	10,177	10,113	10,211	10,168	10,240
d	a	9,795	8,127	8,121	9,734	9,845	9,788	10,343	9,829	9,158	9,024	9,042	8,991
d	b	10,718	10,601	10,970	10,001	11,602	10,258	14,149	11,358	9,461	10,157	9,595	9,986
d	c	10,847	8,558	8,650	10,099	11,769	10,426	14,387	11,424	8,674	8,707	8,608	8,664
d	d	9,945	10,070	10,057	9,778	10,019	9,885	10,510	9,986	9,458	9,412	9,449	9,417

When a cluster sample is taken from frame A, as in Tables 3 and 4, the bias patterns are similar. When there is no misclassification, the MSEs of the optimal and PML estimators are smaller than that of $\hat{Y}(2/3)$ because they account for the survey design. With misclassification, though, the MSE advantage is reduced because of the increased bias.

To study misclassification with a three-frame survey, we selected simple random samples from each frame, and had correct classifications for frames B and C. Table 5 shows results for a simulation with three frames and a simple random sample of size 200 from each frame. The population was generated with $N_d = 10,000$ in each domain and domain means $\mu_a = 1, \mu_{ab} = 2, \mu_{ac} = 3, \mu_{abc} = 4, \mu_b = 5, \mu_{bc} = 6, \mu_c = 7$. In this simulation, frames B and C are correctly classified, and the misclassification patterns for frame A are given in the table. We also studied other domain means, population domain sizes, and sample sizes using a factorial design; results for the other settings showed a similar pattern and are not shown here. The multiplicity estimator \hat{Y}_{ave} , with $m_i = 1$ for $i \in \{a, b, c\}$, $m_i = 1/2$ for $i \in \{ab, ac, bc\}$, and $m_i = 1/3$ for $i \in abc$, is optimal when there is no misclassification, and it equals the unraked single frame estimator. The other fixed weight estimators studied are $\hat{Y}_{2/3}$, with $m^{(A,a)} = m^{(B,b)} = m^{(C,c)} = 1$, $m^{(A,ab)} = m^{(A,ac)} = m^{(A,abc)} = 2/3$, $m^{(B,ab)} = m^{(C,ac)} = 1/3$, and $m^{(B,abc)} = m^{(C,abc)} = 1/6$, and the screening estimator \hat{Y}_{scr} , with $m^{(A,a)} = m^{(B,b)} = m^{(C,c)} = m^{(A,ab)} = m^{(A,ac)} = m^{(A,abc)} = m^{(B,bc)} = 1$.

As with the two-frame study, the bias-corrected estimators are approximately unbiased. The screening estimator is also approximately unbiased since only $\mathcal{S}(A)$ is misclassified. The other estimators all exhibit substantial bias with at least some of the misclassification patterns. For the simulation settings in Table 5, the poststratified, single frame raking, Hartley, and PML estimators exhibit large bias but nevertheless have smaller mean squared error than the fixed weight and bias-corrected estimators; this MSE ordering does not hold in some of the other simulation settings.

Mecatti (2007) and Rao and Wu (2010) argued that the fixed weight multiplicity estimator \hat{Y}_{ave} is unbiased if the only misclassification is among domains that belong to the same number of frames. Misclassifying observations from domain ab to domain ac (pattern c) results in no bias because the weight adjustment in both domains is $1/2$. In practice, though, one would expect pattern (c), with two errors in domain membership (not reporting membership in frame B and erroneously reporting membership in frame C), to be less likely to occur in practice than misclassifying an observation in ab as either a or abc ; \hat{Y}_{ave} can be very sensitive to the latter forms of misclassification. Although a fixed weight estimator is insensitive to misclassification among domains in which the weight adjustments are equal, in these simulations every fixed weight estimator exhibits significant bias for at least some misclassification patterns.

Tables 1 to 5 show that each estimator from Section 2 can exhibit severe bias from domain misclassification. We

recommend that the possible extent of domain misclassification be studied during the survey pretesting phase, so that this information can be used in the survey design. If misclassification probabilities are known accurately, then it may be possible to choose a fixed weight estimator that is insensitive to the presumed form of misclassification. When a misclassification-robust estimator cannot be found or when it is inefficient, the fixed weight estimators can be adjusted to reduce the bias. It should be noted that the bias-corrected weights proposed in Section 6.1 are sensitive to

the input misclassification probabilities. They also do not account for other nonsampling errors such as nonresponse; applying the misclassification weight adjustments in Section 6.1 followed by the nonresponse weight adjustments described in Brick *et al.* (2011) may result in final weights that correct neither for misclassification nor for nonresponse. If domain misclassification and nonresponse are both present, weight adjustments are needed that deal with both problems simultaneously.

Table 3
Estimated bias for dual frame misclassification, with $n_A = 200$, $n_B = 100$, a cluster sample taken from frame A and a simple random sample taken from frame B. MPA and MPB refer to the misclassification patterns for frames A and B

MPA	MPB	$\hat{Y}(1/2)$	$\hat{Y}(1/2)_{\text{post1}}$	$\hat{Y}(1/2)_{\text{post2}}$	$\hat{Y}(1/2)_{bc}$	$\hat{Y}(2/3)$	$\hat{Y}(2/3)_{bc}$	$\hat{Y}(1)$	\hat{Y}_H	\hat{Y}_{FB}	\hat{Y}_{PML}	\hat{Y}_{PEL}	$\hat{Y}_{SF, rake}$
a	a	-148	-142	-139	-148	-155	-155	-170	-312	63	-119	-172	-184
a	b	-5,090	4,199	4,599	-72	-6,774	-84	-10,144	-4,976	1,210	3,615	2,181	1,025
a	c	-5,069	-1,088	-851	-72	-6,759	-96	-10,139	-4,800	-1,994	177	-1,136	-3,216
a	d	-39	-8,379	-8,383	-35	-63	-58	-111	-237	-5,757	-5,909	-5,961	-6,996
b	a	1,168	-1,221	-1,258	-79	768	-63	-32	1,395	-1,690	-1,663	-2,514	-3,170
b	b	-3,716	2,979	3,236	60	-5,784	79	-9,918	-2,815	-86	1,776	-346	-2,087
b	c	-3,704	-2,108	-2,074	73	-5,771	92	-9,905	-2,561	-2,970	-1,410	-3,267	-5,814
b	d	1,317	-9,455	-9,610	95	926	123	144	1,609	-7,285	-7,317	-7,938	-9,498
c	a	1,179	1,281	1,304	-66	772	-58	-41	1,486	1,831	1,652	943	840
c	b	-3,879	5,545	6,087	-118	-5,971	-126	-10,156	-2,972	3,532	4,597	2,405	1,683
c	c	-3,811	318	636	-44	-5,893	-42	-10,058	-2,671	110	1,128	-784	-2,328
c	d	1,423	-6,858	-6,973	191	1,022	206	220	1,824	-4,328	-4,014	-4,516	-5,624
d	a	-33	2,282	2,290	-28	-35	-32	-40	-148	3,627	3,138	3,103	3,728
d	b	-4,974	6,514	7,123	46	-6,660	30	-10,033	-4,863	4,768	6,274	4,742	4,549
d	c	-4,951	1,412	1,883	80	-6,621	84	-9,961	-4,682	1,357	2,863	1,451	388
d	d	42	-5,987	-5,991	53	40	52	37	-126	-2,899	-2,780	-2,791	-3,317

Table 4
Estimated MSE for dual frame misclassification, with $n_A = 200$, $n_B = 100$, a cluster sample taken from frame A and a simple random sample taken from frame B. MPA and MPB refer to the misclassification patterns for frames A and B

MPA	MPB	$\hat{Y}(1/2)$	$\hat{Y}(1/2)_{\text{post1}}$	$\hat{Y}(1/2)_{\text{post2}}$	$\hat{Y}(1/2)_{bc}$	$\hat{Y}(2/3)$	$\hat{Y}(2/3)_{bc}$	$\hat{Y}(1)$	\hat{Y}_H	\hat{Y}_{FB}	\hat{Y}_{PML}	\hat{Y}_{PEL}	$\hat{Y}_{SF, rake}$
a	a	10,916	8,912	8,899	10,916	11,092	11,092	11,879	10,975	9,250	9,155	10,109	9,418
a	b	11,786	10,186	10,324	11,157	12,743	11,503	15,463	12,253	8,906	9,391	10,123	9,231
a	c	11,983	9,575	9,537	11,409	12,922	11,814	15,600	12,395	9,575	9,279	10,391	10,039
a	d	11,042	12,357	12,375	10,941	11,250	11,173	12,056	11,051	11,591	11,605	12,229	12,053
b	a	10,698	9,133	9,154	10,872	10,921	11,049	11,875	10,823	9,255	9,151	10,195	9,766
b	b	10,957	9,803	9,867	11,071	12,033	11,413	15,262	11,215	8,681	8,748	9,610	9,182
b	c	11,115	9,860	9,846	11,272	12,172	11,675	15,361	11,306	9,721	9,252	10,558	10,970
b	d	10,988	13,269	13,408	11,046	11,222	11,262	12,143	11,084	12,484	12,347	13,279	13,598
c	a	10,995	9,090	9,073	11,106	11,187	11,254	12,028	11,125	9,309	9,190	9,798	9,389
c	b	11,104	10,779	11,015	11,090	12,162	11,380	15,348	11,430	9,450	9,724	9,754	9,144
c	c	11,155	9,425	9,400	11,189	12,234	11,600	15,424	11,389	9,219	9,064	9,868	9,658
c	d	10,922	11,328	11,421	10,896	11,121	11,091	11,929	11,017	10,759	10,456	11,151	11,182
d	a	11,011	9,080	9,045	10,920	11,181	11,103	11,913	11,041	9,873	9,579	10,375	10,135
d	b	11,838	11,357	11,669	11,164	12,723	11,453	15,299	12,337	10,258	10,848	11,009	10,403
d	c	11,804	9,334	9,371	11,159	12,707	11,548	15,298	12,224	9,349	9,507	10,102	9,442
d	d	11,179	10,839	10,854	10,989	11,355	11,199	12,059	11,195	10,440	10,302	10,916	10,519

Table 5
Estimated bias and MSE for misclassification in a 3-frame survey, with $n_A = n_B = n_C = 200$ and a simple random sample taken from each frame. MPA refers to the misclassification patterns for frame A. Pattern (a) has no misclassification; (b) $\phi_{aa}^A = 0.8$, $\phi_{a,ab}^A = 0.1$, $\phi_{a,abc}^A = 0.1$, $\phi_{ab,ab}^A = 1$, $\phi_{ac,ac}^A = 1$, $\phi_{abc,abc}^A = 1$; (c) $\phi_{aa}^A = 1$, $\phi_{ab,ab}^A = 0.9$, $\phi_{ab,ac}^A = 0.1$, $\phi_{ac,ac}^A = 1$, $\phi_{abc,abc}^A = 1$; (d) $\phi_{aa}^A = 1$, $\phi_{ab,ab}^A = 0.9$, $\phi_{ab,abc}^A = 0.1$, $\phi_{ac,ac}^A = 1$, $\phi_{abc,abc}^A = 1$; (e) $\phi_{aa}^A = 1$, $\phi_{ab,ab}^A = 0.8$, $\phi_{ab,a}^A = 0.1$, $\phi_{ab,abc}^A = 0.1$, $\phi_{ac,ac}^A = 1$, $\phi_{abc,abc}^A = 1$

	MPA	\hat{Y}_{ave}	$\hat{Y}_{ave, post1}$	$\hat{Y}_{ave, post2}$	$\hat{Y}_{ave, bc}$	$\hat{Y}_{2/3}$	$\hat{Y}_{2/3, bc}$	\hat{Y}_{scr}	\hat{Y}_H	\hat{Y}_{PML}	$\hat{Y}_{SF, rake}$
bias	a	-8	31	28	-8	5	5	20	9	-26	-208
	b	-938	-1,409	-1,478	57	-586	77	107	-2,039	-5,676	-5,624
	c	-26	-485	-508	-26	6	6	6	-324	-825	-957
	d	-231	-514	-557	104	108	108	85	-326	-1,321	-1,438
	e	704	287	247	34	697	27	-4	1,488	1,420	1,193
MSE	a	9,003	4,419	4,410	9,003	10,013	10,013	13,108	7,990	7,281	7,293
	b	8,961	4,711	4,730	8,955	9,952	9,953	13,092	8,085	9,107	9,074
	c	9,119	4,432	4,422	9,119	10,140	10,140	13,238	8,112	7,396	7,422
	d	8,894	4,405	4,405	8,893	9,874	9,874	12,919	7,957	7,414	7,433
	e	9,088	4,438	4,424	9,059	10,071	10,046	13,180	8,254	7,621	7,581

7. Design issues

As discussed in Section 1, multiple frame designs can give better coverage and precision than a single frame survey with equivalent cost. The design problem is more complex than with a single frame survey, though, since a design that is optimal for frame A and frame B separately may not be optimal for the combined sample. Similarly, a design that is optimal when estimator $\hat{Y}(1/2)$ is used may not be optimal for \hat{Y}_{PML} .

Hartley (1962, 1974) derived optimal designs for the estimator $\hat{Y}(\hat{\theta}_H)$ when a simple random sample is taken in each frame. The optimal sample sizes n_A and n_B depend on the relative costs of sampling from the two frames, and on the means and variances of the response variable within the domains. Cochran (1977, pages 144-145) described the dual frame survey in Figure 1 in his chapter on stratified sampling. In this situation, N_a and N_{ab} may be known, especially if frame B is a list frame. Domains a and ab are treated as strata; there is one sample from stratum a and two independent samples from stratum ab . The design problem may be approached as a stratified sample design.

In general, the optimal design is a function of sampling variances and nonsampling errors in each frame, as well as of the estimator chosen. Biemer (1984) and Lepkowski and Groves (1986) discussed designs for the situation in Figure 1 when a stratified multistage sample is taken from each frame, using the Hartley estimator $\hat{Y}(\hat{\theta}_H)$. Lepkowski and Groves (1986) considered interviewer variability and mode bias as well as sampling error when assessing the precision of various designs; frames with less mode bias are allotted higher sample sizes. Brick (2010) derived optimal allocations in the presence of nonresponse, and found that considering the nonresponse when allocating resources to the two frames can greatly increase efficiency in both screening and overlap dual frame surveys.

One of the advantages of a multiple frame design is its flexibility; it is well suited for a modular approach to survey design. In some situations, it may be practical to take an initial sample from the general population (frame A in Figure 4). The design of the samples from frames B and C, corresponding to subpopulations of interest, can then be determined using information in the frame-A sample. For example, if the frame-A sample yields too few engineers, the sample size from an engineering society membership list frame can be correspondingly increased.

Rao (2003) suggested using multiple frame surveys to improve the accuracy of small area estimates in subgroups of interest. In this application, supplemental surveys can be taken in frames with high concentrations of subgroups of interest. As research needs change, resources can be re-allocated among the supplemental surveys without changing the main survey design. A crime victimization survey that uses a national area frame may be supplemented by local victimization surveys; as victimization patterns change, the local surveys can have different sample sizes or be moved to other geographic regions.

Most survey designs attempt to achieve efficiency for the important responses, but in some situations a design that is efficient for one response is inefficient for others. For a survey in which each of four responses of interest was highly correlated with one of the possible stratification variables (but not necessarily correlated with the other stratification variables), Skinner, Holmes and Holt (1994) used a multiple frame survey with four independent stratified samples drawn from a common sampling frame. Each sample was stratified using the stratification variable that was correlated with one of the responses of interest, and so was highly efficient for that response. In estimation, information from all four samples was combined.

Multiple frame surveys can also be used in conjunction with sequential or adaptive sampling methods to improve

yield of a rare or hard-to-reach population such as recent immigrants. For example, a stratified multistage sampling design might be employed for frame A, while an adaptive cluster sampling design (Thompson 2002) might be used for frame B. Domain estimates can be calculated separately for the two designs, and then combined using methods in Section 2. In this situation, frames A and B may completely overlap, so that domain misclassification will not be an issue.

8. Conclusions

In this paper, we have summarized some of the issues involved in using multiple frame methods for U.S. household surveys. Multiple frame designs have great potential for improving efficiency of data collection in household surveys. They can improve coverage by combining incomplete frames, improve the accuracy of estimates for subgroups or rare populations, and increase the flexibility and responsiveness of federal data collection. Multiple frame surveys can facilitate sampling hard-to-reach populations such as recent immigrants or households with infants; a general population survey can be combined with an adaptive sample design or a list frame of births.

In many cases, multiple frame surveys can provide more accurate estimates of population quantities without increasing data collection costs, but the design and estimator must be chosen carefully to realize these savings. A multiple frame survey, like other surveys, may have nonresponse, mode effects, and measurement errors. In addition, unless all of the frames consist of the entire population, multiple frame survey estimators can be sensitive to domain misclassification. One correction for misclassification was given in Section 6, but more research is needed on these challenges. Effects of domain misclassification, nonresponse, and mode bias may be confounded. A designed experiment may help disentangle these effects. We are currently studying the relation among these three types of nonsampling errors. Each form of nonsampling error affects the accuracy of multiple frame estimators, and anticipated nonsampling errors need to be incorporated in an optimal design.

Acknowledgements

This research was partially supported by the National Science Foundation under grants SES-0604373 and DRL-0909630. Some of the material in this paper was presented at the 2010 annual meeting of the Italian Statistical Society and published in the proceedings for that conference. The

author is grateful to the reviewers for their helpful comments.

References

- Bankier, M.D. (1986). Estimators based on several stratified samples with applications to multiple frame surveys. *Journal of the American Statistical Association*, 81, 1074-1079.
- Biemer, P.P. (1984). Methodology for optimal dual frame sample design. Bureau of the Census SRD Research Report CENSUS/SRD/RR-84/07.
- Brick, J.M. (2010). Dual frame landline and cell phone surveys. Paper presented at the annual meeting of the Statistical Society of Canada, Québec City.
- Brick, J.M., Cervantes, I.F., Lee, S. and Norman, G. (2011). Nonsampling errors in dual frame telephone surveys. *Survey Methodology*, 37, 1, 1-12.
- Brick, J.M., Dipko, S., Presser, S., Tucker, C. and Yuan, Y. (2006). Nonresponse bias in a dual frame survey of cell and landline numbers. *Public Opinion Quarterly*, 70, 780-793.
- Chambers, R., Chipperfield, J., Davis, W. and Kovačević, M. (2008). Inference based on estimating equations and probability-linked data. University of Wollongong Centre for Statistical & Survey Methodology Working Paper 18-09.
- Citro, C.F., and Kalton, G., Eds. (2007). *Using the American Community Survey: Benefits and Challenges*. Washington, D.C.: National Academies Press.
- Clark, J., Winglee, M. and Liu, B. (2007). Handling imperfect overlap determination in a dual-frame survey. *Proceedings of the Survey Research Methods Section*, American Statistical Association, 3233-3238.
- Cochran, W.G. (1977). *Sampling Techniques*, 3rd Ed. New York: John Wiley & Sons, Inc.
- de Leeuw, E. (2008). Choosing the method of data collection. In *International Handbook of Survey Methodology*, (Eds., E. de Leeuw, J. Hox and D. Dillman). New York: Lawrence Erlbaum, 113-135.
- de Leeuw, E., Hox, J. and Dillman, D. (2008). Mixed-mode surveys: When and why. In *International Handbook of Survey Methodology*, (Eds., E. de Leeuw, J. Hox and D. Dillman). New York: Lawrence Erlbaum, 299-316.
- Demnati, A., Rao, J.N.K., Hidiroglou, M.A. and Tambay, J.-L. (2007). On the allocation and estimation for dual frame survey data. *Proceedings of the Survey Research Methods Section*, American Statistical Association, 2938-2945.
- Déville, J.-C., and Särndal, C.-E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, 87, 376-382.
- Fuller, W.A., and Burmeister, L.F. (1972). Estimators for samples selected from two overlapping frames. *Proceedings of the Social Statistics Section*, American Statistical Association, 245-249.

- González-Villalobos, A., and Wallace, M.A. (1996). *Multiple Frame Agriculture Surveys*, Rome: Food and Agriculture Organization of the United Nations. Vols. 1 and 2.
- Haines, D.E., and Pollock, K.H. (1998). Combining multiple frames to estimate population size and totals. *Survey Methodology*, 24, 79-88.
- Hansen, M.H., Hurwitz, W.N. and Madow, W.G. (1953). *Sample Survey Methods and Theory*. New York: John Wiley & Sons, Inc. Volume 1.
- Hartley, H.O. (1962). Multiple frame surveys. *Proceedings of the Social Statistics Section*, American Statistical Association, 203-206.
- Hartley, H.O. (1974). Multiple frame methodology and selected applications. *Sankhyā*, Series C, 36, 99-118.
- Horvitz, D.G., and Thompson, D.J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47, 663-685.
- Iachan, R., and Dennis, M.L. (1993). A multiple frame approach to sampling the homeless and transient population. *Journal of Official Statistics*, 9, 747-764.
- Kalton, G., and Anderson, D.W. (1986). Sampling rare populations. *Journal of the Royal Statistical Society*, Series A, 149, 65-82.
- Kennedy, C. (2007). Evaluating the effects of screening for telephone service in dual frame RDD surveys. *Public Opinion Quarterly*, 71, 750-771.
- Kott, P.S., Amrhein, J.F. and Hicks, S.D. (1998). Sampling and estimation from multiple list frames. *Survey Methodology*, 24, 3-9.
- Lepkowski, J.M., and Groves, R.M. (1986). A mean squared error model for multiple frame, mixed mode survey design. *Journal of the American Statistical Association*, 81, 930-937.
- Lesser, V.M., and Kalsbeek, W.D. (1999). Nonsampling errors in environmental surveys. *Journal of Agricultural, Biological, and Environmental Statistics*, 4, 473-488.
- Lohr, S.L. (2007). Recent developments in multiple frame surveys. *Proceedings of the Survey Research Methods Section*, American Statistical Association, 3257-3264.
- Lohr, S.L. (2009). Multiple frame surveys. In *Handbook of Statistics, Sample Surveys: Design, Methods and Applications*, (Eds., D. Pfeffermann and C.R. Rao). Amsterdam: North Holland, Vol. 29A, 71-88.
- Lohr, S.L., and Rao, J.N.K. (2000). Inference in dual frame surveys. *Journal of the American Statistical Association*, 95, 271-280.
- Lohr, S.L., and Rao, J.N.K. (2006). Estimation in multiple-frame surveys. *Journal of the American Statistical Association*, 101, 1019-1030.
- Lu, W., Brick, J.M. and Sitter, R. (2006). Algorithms for constructing combined strata variance estimators. *Journal of the American Statistical Association*, 101, 1680-1692.
- Lu, Y., and Lohr, S.L. (2010). Gross flow estimation in dual frame surveys. *Survey Methodology*, 36, 13-22.
- Mecatti, F. (2007). A single frame multiplicity estimator for multiple frame surveys. *Survey Methodology*, 33, 151-157.
- National Science Foundation (2003). *SESTAT: Design and Methodology*, <http://srsstats.sbe.nsf.gov/docs/techinfo.html>.
- Rao, J.N.K. (2003). *Small Area Estimation*. New York: John Wiley & Sons, Inc.
- Rao, J.N.K., and Wu, C. (2010). Pseudo-empirical likelihood inference for dual frame surveys. *Journal of the American Statistical Association*, 105, 1494-1503.
- Skinner, C.J. (1991). On the efficiency of raking ratio estimation for multiple frame surveys. *Journal of the American Statistical Association*, 86, 779-784.
- Skinner, C.J., Holmes, D.J. and Holt, D. (1994). Multiple frame sampling for multivariate stratification. *International Statistical Review*, 62, 333-347.
- Skinner, C.J., and Rao, J.N.K. (1996). Estimation in dual frame surveys with complex designs. *Journal of the American Statistical Association*, 91, 349-356.
- Thompson, S.K. (2002). *Sampling Techniques*, 2nd Ed. New York: John Wiley & Sons, Inc.
- Vannieuwenhuyze, J., Loosveldt, G. and Molenberghs, G. (2011). A method for evaluating mode effects in mixed-mode surveys. *Public Opinion Quarterly*, 74, 1027-1045.

Ten years of balanced sampling with the cube method: An appraisal

Yves Tillé¹

Abstract

This paper presents a review and assessment of the use of balanced sampling by means of the cube method. After defining the notion of balanced sample and balanced sampling, a short history of the concept of balancing is presented. The theory of the cube method is briefly presented. Emphasis is placed on the practical problems posed by balanced sampling: the interest of the method with respect to other sampling methods and calibration, the field of application, the accuracy of balancing, the choice of auxiliary variables and ways to implement the method.

Key Words: Sampling; Balancing; Horvitz-Thompson estimator.

1. Introduction

While the idea of balanced sampling has been around since the early days of survey statistic development, applying the concept has been difficult because almost all the proposed methods have either been enumerative or rejective and required considerable computation time. The algorithm of the cube method was proposed in 1998 by Deville and Tillé, and a first implementation was written by three students of the École Nationale de la Statistique et de l'Analyse de l'Information of Rennes in France (see Bousabaa, Lieber and Sirolli 1999). Finally, the method was published in Tillé (2001) and Deville and Tillé (2004). Since this time, several implementations have been proposed and several survey managers have used the cube method to select samples, the most important applications being the New French Census and the French Master Sample.

Our aim is to assess 10 years of development and use of balanced sampling in order to better ascertain when and how the cube method can be used to select samples of householders or establishments. After discussing the concept of balanced sample and balanced sampling in Section 2, we give a list of particular cases in Section 3. In Section 4, we briefly trace the history of this concept for both the model-based and design-based frameworks. Next, in Section 5, we provide a brief overview of the cube method, which is a class of algorithms that allows us to select randomly balanced samples with given inclusion probabilities (see Deville and Tillé 2004; Tillé 2001, 2006b). We try to present the main principles of this algorithm without giving a detailed description of the technicalities of the method. Section 6 is devoted to the principles of variance estimation in balanced sampling. Finally, in Sections 7, we discuss the interest of balanced sampling in practice and compare balanced sampling with other sampling methods and calibration. We also give a list of recent applications. This Section also deals with the accuracy of balancing, the

choice of auxiliary variables and ways to implement balanced sampling. The paper ends with an exhaustive bibliographical list of references on balanced sampling and their applications.

2. Balanced sampling

2.1 Definition of a balanced sample

Consider a sample s of size n that is a subset of a finite population U of size N . A sample is said to be balanced if, for a vector of auxiliary variable $\mathbf{x}_k = (x_{k1}, \dots, x_{kj}, \dots, x_{kp})'$,

$$\frac{1}{n} \sum_{k \in s} \mathbf{x}_k = \frac{1}{N} \sum_{k \in U} \mathbf{x}_k, \quad (1)$$

which means that the sample means of the x -variables match their population means.

Brewer (1999) drew a distinction between a balanced selection of samples and a random selection of samples. However, a balanced sample may be selected randomly. If a random sample S is selected randomly, then each unit of the population has an inclusion probability π_k of being selected. In this case, a random sample must satisfy the following balancing equations:

$$\sum_{k \in S} \frac{\mathbf{x}_k}{\pi_k} = \sum_{k \in U} \mathbf{x}_k. \quad (2)$$

In other words, in a balanced sample, the total of the x -variables are estimated without error. Several authors like Cumberland and Royall (1981) and Kott (1986) would call a sample that satisfies Equation (2) a ' π -balanced sample', as opposed to a 'mean-balanced sample' defined by Equation (1). Nevertheless, in this paper, we will consider that (1) is only a particular case of (2) that occurs when $\pi_k = n/N$ or when the sample is not selected randomly. We refer to both cases as a balanced sample.

1. Yves Tillé, University of Neuchâtel, Pierre à Mazel 7, 2000 Neuchâtel Switzerland. E-mail : yves.tille@unine.ch.

2.2 Balanced sampling design

Let $p(s)$ denote the sampling design, *i.e.*, the probability that sample s is selected, such that $p(s) = \Pr(S = s)$, where S is the random sample and $n(S)$ the size of the sample S . According to the definition of Deville and Tillé (2004), a sampling design $p(\cdot)$ is said to be *balanced* on auxiliary variables x_1, \dots, x_p if the Horvitz-Thompson estimator satisfies Equation (2). In a balanced sampling design, the inclusion probabilities are decided prior to sampling. A balanced sampling can be viewed as a kind of calibration that is directly integrated into the sampling design. The main problem is that the balancing equations (2) can rarely be exactly satisfied. We refer to this difficulty as the 'rounding problem'.

Example 1. If $N = 4$, $n = 2$, $\pi_k = 1/2$, for all $k \in U$ and $x_1 = 0$, $x_2 = 1$, $x_3 = 2$, $x_4 = 4$, then the balancing equations given in (2) becomes

$$\frac{1}{n} \sum_{k \in s} x_k = \frac{1}{N} \sum_{k \in U} x_k,$$

which is equivalent to

$$\sum_{k \in s} x_k = \frac{n}{N} \sum_{k \in U} x_k. \quad (3)$$

Since

$$\frac{n}{N} \sum_{k \in U} x_k = \frac{2}{4} (0 + 1 + 2 + 4) = 3.5,$$

and the left side of (3) is always an integer, then an exactly balanced sample does not exist.

Indeed, sample selection is an integer problem. The cube method therefore aims to select a sample that exactly satisfies the inclusion probabilities π_k while remaining as balanced as possible.

3. Special cases of balanced sampling

3.1 Unequal probability sampling and stratification

Some well-known sampling designs are particular cases of balanced sampling:

1. Sampling with a fixed sample size is a particular case of balanced sampling. In this case, the only balancing variable is π_k . The balancing equations given in (2) become

$$\sum_{k \in s} \frac{\pi_k}{\pi_k} = \sum_{k \in S} 1 = \sum_{k \in U} \pi_k,$$

which means that the sample size must be fixed.

2. Stratification is a particular case of balanced sampling. Suppose that the population is partitioned in H strata U_h , $h = 1, \dots, H$, of sizes N_h , $h = 1, \dots, H$, and that a sample is selected in each stratum by

means of simple random sampling without replacement with fixed sample size n_h , $h = 1, \dots, H$. In this case, the balancing variables are the indicator variables of the strata

$$\delta_{kh} = \begin{cases} 1 & \text{if } k \in U_h \\ 0 & \text{otherwise.} \end{cases}$$

Under a stratified design, the Horvitz-Thompson estimators of the sizes of the strata exactly equal the sizes of the strata, which is a property of balancing on the indicator variables of the strata. Indeed, since the inclusion probabilities in stratum h are $\pi_k = n_h / N_h$, $k \in U_h$, the balancing equations become

$$\sum_{k \in s} \frac{N_h \delta_{kh}}{n_h} = \sum_{k \in U} \delta_{kh} = N_h, \quad h = 1, \dots, H,$$

and are exactly satisfied.

These two designs are well known and are commonly applied in official statistics in order to reduce variance. The more general concept of balancing allows more freedom to choose the most appropriate balancing variables that will improve the accuracy of the estimators.

3.2 Overlapping strata

Constructing a stratified sampling design is often a difficult exercise. Statisticians often try to stratify using several qualitative variables. However, in most cases, crossing all of the strata of all the variables will cause the cells to become too small for a sample to be selected in each cell. In the context of calibration, statisticians generally calibrate on marginal totals and not on all the cells contained in a contingency table. Since a balanced sampling can be viewed as a kind of calibration that is directly integrated in the sampling design, one would also like to balance using only marginal totals. Nevertheless, the usual theory of stratification does not allow overlapping strata since the stratification must be a partition of the population. Now, the cube method enables us to directly balance on totals of overlapping strata by simply using the indicators of the strata as balancing variables.

3.3 Balancing on a constant

Another interesting special case of balanced sampling occurs when a constant is used as a balancing variable. If $\mathbf{x}_k = 1$ for all $k \in U$, the balancing equations become

$$\sum_{k \in s} \frac{1}{\pi_k} = \sum_{k \in U} 1 = N.$$

Actually,

$$\sum_{k \in S} \frac{1}{\pi_k}$$

is the Horvitz-Thompson estimator of N . This means that, if a constant is used as a balancing variable, the estimated population size matches the known size N , which is far from being a given when the statistical units are selected with unequal inclusion probabilities.

4. History of the concept of balancing and existing methods

The idea of balanced sampling is very old and is linked to the vague concept of representativeness that was already used by Kiaer (1896, 1899, 1903, 1905). The first paper dedicated to the selection of a balanced sample is due to Gini (1928) and Gini and Galvani (1929) who selected a sample of 29 from the 214 Italian districts in order to match several population totals. Both Neyman (1952) and Yates (1960) condemned the paper of Gini and Galvani essentially because this sample was not randomly selected (see Langel and Tillé 2010). The first methods for selecting a random balanced sample were proposed by Yates (1946) and Thionet (1953), but these methods were rejective in the sense that they involved selecting samples or changing units randomly in the sample until a balanced enough sample was obtained.

In the model-based framework, Royall (1976a, b) advocated the use of balanced sampling in order to reach the optimal strategy and to protect against mis-specification of the model. (see also Royall and Pfeffermann 1982; Kott 1986; Cumberland and Royall 1988; Royall 1988; Tirari 2006; Nedyalkova and Tillé 2009). While several methods for selecting a balanced sample are presented in the book of Valliant, Dorfman and Royall (2000), these methods do not necessarily specify the inclusion probabilities of the sample. In the model-based framework, it is important to have a balanced sample. However, this sample does not always need to be randomly selected.

Hájek (1981) also advocated the use of balanced sampling. For Hájek, a balanced sampling is a particular case of representative strategy, a strategy being a couple made of a sampling design and an estimator. A representative strategy is a strategy that estimates the totals of auxiliary variables without error. In this sense, a balanced sampling design with the Horvitz-Thompson estimator is a representative strategy. Hájek (1981) proposes a rejective procedure that consists of selecting a sequence of samples until a balanced one is obtained. Rejective procedures have two drawbacks: if several balancing variables are used, the procedure can be very slow; secondly, the inclusion probabilities of rejective designs are not the same as the original design. The inclusion probabilities of statistical units that are close to the population means are increased to the detriment of the units

that are far from the center (see for instance the simulations of Legg and Yu 2010).

Another method of selection consists of enumerating all the possible samples, and then constructing a sampling design only to select the samples that are adequately balanced. Such a design can be constructed by using linear programming. This technique was applied by Ardilly (1991) to select the primary units of the French master sample. Nevertheless, this method can only be applied on small population sizes because of the combinatory explosion of the number of samples when the size of the population is large.

Deville, Grosbras and Roth (1988) and Deville (1992) proposed multivariate methods for balanced sampling with equal inclusion probabilities. Hedayat and Majumdar (1995) have proposed the adaptation of an experimental design technique that would enable a balanced sampling design to be constructed. Again, this technique is restricted to equal inclusion probabilities. Finally, the cube method was proposed by Deville and Tillé (2004). This method is general in the sense that the inclusion probabilities are exactly satisfied, that these probabilities may be equal or unequal and that the sample is as balanced as possible.

Fuller (2009) studied a rejective procedure by fixing a tolerance interval outside of which the sample is rejected and proposed an estimator of variance. Even if the inclusion probabilities are changed with a rejective procedure, Fuller (2009) shows that efficient estimates are obtained by using the inclusion probabilities of the original design. Using a set of simulations, Legg and Yu (2010) compared this rejective procedure to the cube method and showed that both methods perform equally. Finally, Dudoignon and Vanheuverzwyn (2006) proposed a fast method of balanced sampling for marginal totals, whereas Périé (2008) proposed a method based on permanent random numbers that provides a balanced sample. With the Périé (2008) method, the inclusion probabilities are only approximately satisfied.

5. The cube method

5.1 Main ideas

The cube method (see Deville and Tillé 2004; Tillé 2001, 2006a, b; Ardilly 2006) is a class of sampling algorithms that selects a balanced sample and exactly satisfies a set of given inclusion probabilities. The cube method is an extension of the splitting method that was developed by Deville and Tillé (1998). It is based on a random transformation of the vector of inclusion probabilities until a sample is obtained such that:

- (i) the inclusion probabilities are exactly satisfied,
- (ii) the balancing equations are satisfied to the furthest extent possible.

The name of the method comes from the geometric representation of a sampling design. Indeed, a sample may be represented by a vector of samples indicators:

$$\mathbf{s} = (I[1 \in s] \dots I[k \in s] \dots I[N \in s])',$$

where $I[k \in s]$ takes value 1 if $k \in s$ and 0 if not. A sample may thus be viewed as a vertex of an N -cube as showed in Figure 1.

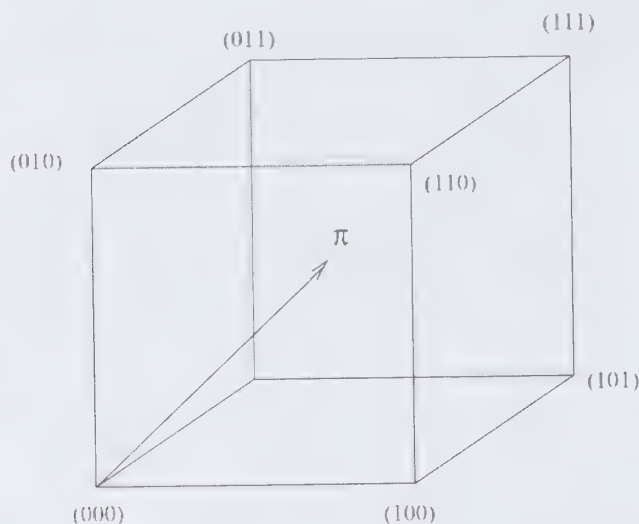


Figure 1 Possible samples in a population of size $N = 3$

Let us also define

$$E(\mathbf{s}) = \sum_{s \in S} p(\mathbf{s}) \mathbf{s} = \boldsymbol{\pi},$$

where $\boldsymbol{\pi} = [\pi_k]$ is the vector of inclusion probabilities. The balancing equations

$$\sum_{k \in S} \frac{\mathbf{x}_k}{\pi_k} = \sum_{k \in U} \mathbf{x}_k,$$

may also be written

$$\sum_{k \in U} \tilde{\mathbf{x}}_k s_k = \sum_{k \in U} \tilde{\mathbf{x}}_k \pi_k, \quad (4)$$

where $s_k \in \{0, 1\}$ and $\tilde{\mathbf{x}}_k = \mathbf{x}_k / \pi_k$, $k \in U$. Expression (4) is a system of equations with unknowns values s_k that define an affine subspace in \mathbb{R}^N of dimension $N - p$ denoted by Q , where

$$Q = \left\{ \mathbf{u} \in \mathbb{R}^N \mid \sum_{k \in U} \tilde{\mathbf{x}}_k u_k = \sum_{k \in U} \tilde{\mathbf{x}}_k \pi_k \right\}.$$

The problem of selecting a balanced sample may thus be reformulated. A balanced sampling design consists of choosing a vertex of the N -cube (a sample) that remains on the linear sub-space Q . Figures 2 and 3 respectively show two examples: the first one is a constraint of fixed sample size and the second one is a constraint that generates a rounding problem.

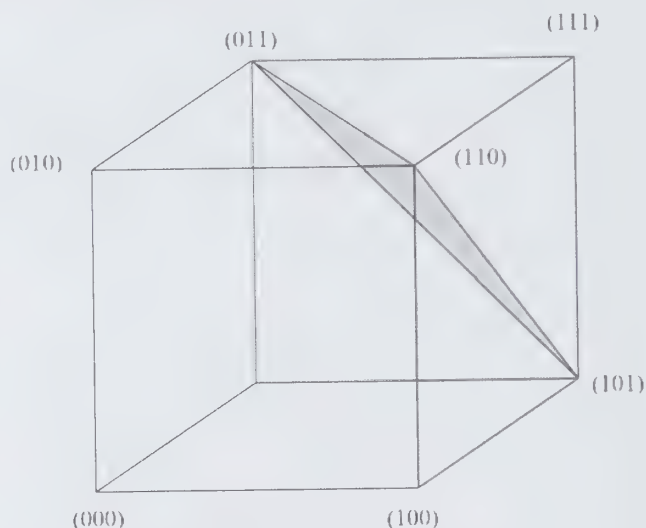


Figure 2 Possible samples in a population of size $N = 3$ with a constraint of fixed sample size $n = 2$

The Cube method (Deville and Tillé 2004) is divided into two phases: the flight phase and the landing phase. The flight phase is a random walk that begins at the vector of inclusion probabilities and remains in the intersection of the cube and the constraint subspace. This random walk stops at a vertex of the intersection of the cube and the constraint subspace. At the end of the flight phase, if a sample is not obtained, the landing phase entails in selecting a sample that is as close as possible to the constraint subspace.

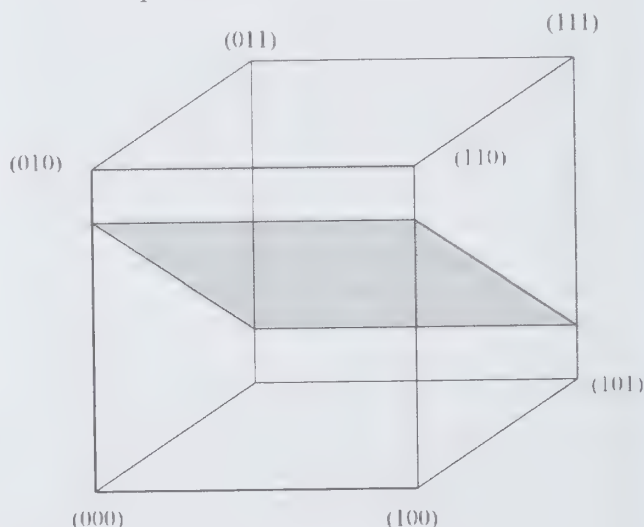


Figure 3 Possible samples in a population of size $N = 3$ with a constraint and a rounding problem

Example 2. If the constraint is the fixed sample size, the flight phase randomly transforms a vector of inclusion probabilities into a vector of 0 and 1. At each step of the algorithm, the vector of inclusion probabilities is transformed randomly, but the sum of inclusion probabilities must remain equal to n . For instance, with $\boldsymbol{\pi} = (0.5, 0.5, 0.5, 0.5)$ and $n = 2$, we are able to obtain the following sequence of vectors:

$$\pi = \begin{pmatrix} 0.5 \\ 0.5 \\ 0.5 \\ 0.5 \end{pmatrix} \rightarrow \begin{pmatrix} 0.6666 \\ 0.6666 \\ 0.6666 \\ 0 \end{pmatrix} \rightarrow \begin{pmatrix} 1 \\ 0.5 \\ 0.5 \\ 0 \end{pmatrix} \rightarrow \begin{pmatrix} 1 \\ 0 \\ 1 \\ 0 \end{pmatrix} = \mathbf{s}.$$

The algorithm ends when all the components of the vector are equal to 0 or 1.

Example 3. If the constraint is the fixed sample size, a rounding problem appears if the sum of inclusion probabilities is not an integer. If there is a rounding problem, then some components cannot be set to zero. For instance, with $\pi = (0.5, 0.5, 0.5, 0.5, 0.5)$ and

$$\sum_{k \in U} \pi_k = 2.5,$$

we may observe the following sequence of vectors:

$$\pi = \begin{pmatrix} 0.5 \\ 0.5 \\ 0.5 \\ 0.5 \\ 0.5 \end{pmatrix} \rightarrow \begin{pmatrix} 0.625 \\ 0 \\ 0.625 \\ 0.625 \\ 0.625 \end{pmatrix} \rightarrow \begin{pmatrix} 0.5 \\ 0 \\ 0.5 \\ 1 \\ 0.5 \end{pmatrix} \rightarrow \begin{pmatrix} 1 \\ 0 \\ 0.25 \\ 1 \\ 0.25 \end{pmatrix} \rightarrow \begin{pmatrix} 1 \\ 0 \\ 0.5 \\ 1 \\ 0 \end{pmatrix} = \pi^*.$$

In this case, the flight phase cannot end with a vector of 0 or 1 of which the sum is equal to 2.5. In this case, the flight phase ends with a vector containing one non-integer component.

5.2 The flight phase

The first step of the flight phase is presented in Figure 4 for a very specific case: the population size $N = 3$. The only balancing constraint is the fixed sample size $n = 2$. At the first step, a vector $\mathbf{u}(0)$ must be chosen. This vector may be chosen freely but must be such that $\pi + \mathbf{u}(0)$ remains in the subspace of constraints. Actually, the cube method is a family of methods that depends on the way the vector $\mathbf{u}(0)$ is chosen. This vector may be chosen randomly or not.

If, from π , we follow the direction given by vector $\mathbf{u}(0)$, then we will necessarily cross a face of the cube. Let us consider this point denoted on Figure 4 by $\pi(0) + \lambda_1^*(0)\mathbf{u}(0)$. Now, if, from π , we follow the opposite direction, i.e., the direction given by vector $-\mathbf{u}(0)$, we will also cross a face of the cube. Let us consider this point denoted on Figure 4 by $\pi(0) - \lambda_2^*(0)\mathbf{u}(0)$. At the first step, vector $\pi(0) = \pi$ is modified randomly. Vector $\pi(1)$ will be set to $\pi(0) + \lambda_1^*(0)\mathbf{u}(0)$ or to $\pi(0) - \lambda_2^*(0)\mathbf{u}(0)$. The choice is done randomly in such a way that $E[\pi(1)] = \pi(0)$. At the end of the first step of the flight phase, we have thus jumped on a face of the cube, which means that at least one component of $\pi(1)$ is equal to 0 or 1, i.e., the problem is reduced from a problem of sampling from a population of size $N = 3$ to a population of size $N = 2$. In N steps at least, the flight phase is thus completed.

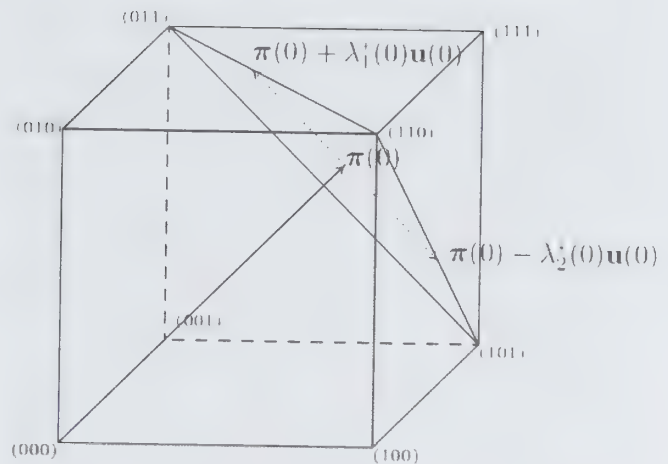


Figure 4 Flight phase in a population of size $N = 3$ with a sample size constraint $n = 2$

More generally, the flight phase is a random walk in the intersection of the balancing subspace and the cube. This random walk stops at a vertex of the intersection of the cube and the subspace. The flight phase is defined by the following class of algorithms. First initialize with $\pi(0) = \pi$. Next, at time $t = 0, \dots, T$,

1. Generate any vector $\mathbf{u}(t) = [u_k(t)] \neq 0$ such that
 - (i) $\mathbf{u}(t)$ is in the kernel of matrix $\mathbf{A} = (\mathbf{x}_1/\pi_1, \dots, \mathbf{x}_k/\pi_k, \dots, \mathbf{x}_N/\pi_N)$, i.e., $\mathbf{A}\mathbf{u}(t) = 0$,
 - (ii) $u_k(t) = 0$ if $\pi_k(t)$ is integer.

2. Compute $\lambda_1^*(t)$ and $\lambda_2^*(t)$, the largest values such that

$$0 \leq \pi(t) + \lambda_1(t)\mathbf{u}(t) \leq 1,$$

$$0 \leq \pi(t) - \lambda_2(t)\mathbf{u}(t) \leq 1.$$

3. Compute

$$\pi(t+1) = \begin{cases} \pi(t) + \lambda_1^*(t)\mathbf{u}(t) & \text{with probability } q_1(t) \\ \pi(t) - \lambda_2^*(t)\mathbf{u}(t) & \text{with probability } q_2(t), \end{cases}$$

$$\text{where } q_1(t) = \lambda_2^*(t) / \{\lambda_1^*(t) + \lambda_2^*(t)\} \text{ and } q_2(t) = 1 - q_1(t).$$

The flight phase stops when it is no longer possible to find a vector $\mathbf{u}(t) \neq 0$.

5.3 Landing phase

If, at the end of the flight phase, the balancing equations are not exactly satisfied, there is a need for a landing phase. Let $\pi^* = [\pi_k^*]$ be the vector obtained at the last step of the flight phase. It is possible to prove (see Deville and Tillé 2004) that

$$\text{card}(U^*) \leq p,$$

where

$$U^* = \{k \in U \mid 0 < \pi_k^* < 1\}$$

and p is the number of balancing variables. The aim of the landing phase is to find a sample \mathbf{s} such that

$E(\mathbf{s}|\pi^*) = \pi^*$, which is almost balanced. There are two ways of selecting such a sample:

1. *The flight phase by linear programming* consists of considering all the possible samples of U^* . A cost is assigned to each sample. This cost, is, for instance, the distance between the sample and the subspace of constraints. Next, one looks for a sampling design on U^* that minimizes the expected cost and that satisfies the inclusion probabilities π^* . This problem can be solved because the number of samples to consider is reasonable due to the small size of U^* .
2. *The flight phase by suppression of variables* may be used when the number of balancing variables is too large for the linear program to be solved by a simplex algorithm ($p > 20$). With this method, an auxiliary variable is dropped at the end of the flight phase. Next, we can return to the flight phase until it is no longer possible to 'move' within the constraint subspace. The constraints are then relaxed successively according to an order of preference.

6. Variance and variance estimation

6.1 A residual technique

The variance of the Horvitz-Thompson estimator can be estimated by using a residual technique developed in Deville and Tillé (2005). The residual technique is comparable to the technique used to estimate the variance of the calibration estimator and has been validated by a set of simulations. The estimated variance of the Horvitz-Thompson estimator is thus very similar to the estimated variance of a generalized regression (GREG) estimator. Nevertheless, the variance of the GREG estimator is generally underestimated because it does not take into account the randomness of the weights. Indeed, if the usual variance of the GREG estimator is computed for the special case of poststratification, we obtain the variance of a stratified design with proportional allocation. The variance of the poststratified estimator is nevertheless larger than the variance in a stratified design with proportional allocation.

6.2 Approximation of variance

If the balanced sampling design has a large entropy, Hájek (1981) and Deville and Tillé (2005, method 4) have proposed the following approximation of the design variance given by:

$$\text{var}_p(\hat{Y}_\pi) \cong \text{var}_{app}(\hat{Y}_\pi) = \sum_{k \in U} d_k \frac{(y_k - \mathbf{x}'_k \mathbf{b})^2}{\pi_k^2}, \quad (5)$$

where the subscript p denotes the sampling design,

$$\mathbf{b} = \left(\sum_{k \in U} d_k \frac{\mathbf{x}_k \mathbf{x}'_k}{\pi_k^2} \right)^{-1} \sum_{k \in U} d_k \frac{\mathbf{x}_k y_k}{\pi_k^2},$$

and the d_k are the solution of the nonlinear system

$$\pi_k (1 - \pi_k) = d_k - \frac{d_k \mathbf{x}'_k}{\pi_k} \left(\sum_{\ell \in U} d_\ell \frac{\mathbf{x}_\ell \mathbf{x}'_\ell}{\pi_\ell^2} \right)^{-1} \frac{d_k \mathbf{x}_k}{\pi_k}, \quad k \in U. \quad (6)$$

The entropy of the sampling design depends on the way vectors $\mathbf{u}(t)$ are chosen during the flight phase. In order to increase the entropy, vector $\mathbf{u}(t)$ can be chosen randomly or the population can be randomly sorted before selecting the sample.

Expression (5), which only uses the first-order inclusion probabilities, was validated by Deville and Tillé (2005) under a variety of balanced samples regardless of how the y -values were generated. An approximation very close to Expression (5) was obtained by Fuller (2009) and Legg and Yu (2010) for a balanced sampling design obtained by a rejective procedure in the case of an initial design that uses Poisson sampling. These approximations do not take the rounding problem into account.

6.3 Estimation of variance

Deville and Tillé (2005) proposed a family of variance estimators for balanced sampling, of the form

$$\widehat{\text{var}}(\hat{Y}_\pi) = \sum_{k \in S} c_k \frac{(y_k - \mathbf{x}'_k \hat{\mathbf{b}})^2}{\pi_k^2}, \quad (7)$$

where

$$\hat{\mathbf{b}} = \left(\sum_{\ell \in S} c_\ell \frac{\mathbf{x}_\ell \mathbf{x}'_\ell}{\pi_\ell^2} \right)^{-1} \sum_{\ell \in S} c_\ell \frac{\mathbf{x}_\ell y_\ell}{\pi_\ell^2}$$

and the c_k are the solutions of the nonlinear system

$$1 - \pi_k = c_k - \frac{c_k \mathbf{x}'_k}{\pi_k} \left(\sum_{\ell \in S} c_\ell \frac{\mathbf{x}_\ell \mathbf{x}'_\ell}{\pi_\ell^2} \right)^{-1} \frac{c_k \mathbf{x}_k}{\pi_k}, \quad (8)$$

which can be solved by a fixed point algorithm.

In Deville and Tillé (2005), simpler variants of c_k were also proposed. For instance, one can use the alternative values,

$$\tilde{c}_k \approx \frac{n}{n-p} (1 - \pi_k),$$

that are very close to c_k . The estimator $\widehat{\text{var}}(\hat{Y}_\pi)$ is approximately design-unbiased because it is an estimator by substitution of the approximation given in expression (5), (for more information regarding estimators obtained by substitution, see Deville 1999), which is a reasonable approximation of the variance under the sampling design.

It is not easy to use bootstrap method to estimate the variance in the context of balanced sampling. Balanced samples with replacement should be selected from the original sample. A generalization of the cube method for balanced sampling with replacement has not yet been described. A solution, proposed by Chauvet (2007), consists of reconstructing an artificial population from the sample.

Next, bootstrap samples are selected by using balanced sampling. Another solution was proposed by Fuller (2010) for balanced rejective sampling. Breidt and Chauvet (2010a) have proposed an alternative method where a martingale difference representation of the cube method is used in order to approximate second-order inclusion probabilities, which enables us to construct a nearly unbiased variance estimator.

7. Balanced sampling in practice

7.1 Interest of balanced sampling

In the model-assisted and the model-based frameworks, a balancing sampling design with the Horvitz-Thompson estimator is often the optimal strategy (see Nedyalkova and Tillé 2009). Indeed, when the sample is balanced, the variances of the Horvitz-Thompson estimators of the auxiliary variables are equal to zero. Under a linear model, the variance of the Horvitz-Thompson estimator of the interest variable will only depend on the residuals of the model.

The advantages of balanced sampling are as follows:

- (i) Balanced sampling increases the accuracy of the Horvitz-Thompson estimator. This point has been developed in Section 6. Indeed, the variance of the Horvitz-Thompson estimator only depends on the residuals of the regression of the interest variable by the balancing variables.
- (ii) Balanced sampling protects against large sampling errors. Indeed, the most unfavourable samples have a null probability of being selected.
- (iii) If the variable of interest is well explained by the auxiliary information, in model-based inference, balanced sampling protects against a mis-specification of the model. This point is largely developed by Royall (1976b, a) and Valliant *et al.* (2000). A recent discussion of this important question is given in Nedyalkova and Tillé (2009, 2010).
- (iv) Balanced sampling can ensure that the sample sizes in planned domains are not too small or - much worse - equal to zero. Indeed, if an indicator variable of the domain is added in the list of auxiliary variables, the size of the domain is then fixed in the sample.
- (v) Balanced sampling allows us to avoid random weights. With balanced sampling, the Horvitz-Thompson weights can be used. If the sampling design does not contain any balancing constraints (for instance with Poisson sampling) the weighting system obtained by a calibration procedure becomes very random, which increases the variance of the estimators. If the sample is balanced, the weights will be less random even if a calibration procedure is used after balancing.

The availability of easy to use packages contributed to the large use of the cube method in several important statistical processes. The first main application of the cube method is selection of the rotation groups for the French census. (See Desplanques 2000; Dumais, Bertrand and Kauffmann 2000; Durr and Dumais 2001, 2002; Dumais and Isnard 2000; Bertrand, Christian, Chauvet and Grosbras 2004; da Silva, da Silva Borges, Aires Leme and Moura Reis Miceli 2006). For the municipalities with fewer than 10,000 inhabitants, five non-overlapping rotation groups of municipalities are selected using a balanced sampling design with equal inclusion probabilities (1/5). Each year, a fifth of the municipalities are surveyed. So after 5 years, all the small municipalities are selected. For the municipalities with more than 10,000 inhabitants, in each municipality, five non-overlapping balanced samples of addresses are selected with inclusion probabilities 8%. So, after 5 years, 40% of the addresses are visited. The balancing variables are socio-demographic variables taken from the last census.

In the French master sample, the primary units are geographical areas that are selected using a balanced sampling design (see Wilms 2000; Christine and Wilms 2003; Christine 2006). The master sample is a self-weighted multi-stage sampling. So the primary units are selected with unequal probabilities that are proportional to their sizes. The balancing variables are socio-demographic variables taken from the last census. Bardaji (2001) and Even (2002) have also used balanced sampling to select a sample of beneficiaries of subsidized jobs. Seven populations are surveyed, a balanced sample of beneficiaries is selected in each of the populations by using between two and five balancing variables according to the populations.

In the company Électricité de France (EDF), new electricity meters allow electricity consumption for each household to be measured on a continuous basis. The amount of information collected is so large that it is impossible to archive all the data. Dessertaine (2006, 2007) used balanced sampling to select the time series of consumption that must be archived in order to ensure that they represent the consumption of the entire French population as accurately as possible. Biggeri and Falorsi (2006) used balanced sampling to improve the quality of the consumer price index in Italy. Gismondi (2007) tested balanced sampling to estimate the number of tourist nights spent in Italy. D'Alò, Di Consiglio, Falorsi and Solari (2006) and Falorsi and Righi (2008) also proposed using a balanced sampling design to estimate totals in small domains. Simulations were run by Marí, Barbará, Mitas and Passamonti (2007b, a) in Argentina and Chipperfield (2009) in Australia to assess the interest of balanced sampling for the master sample.

At Statistics Canada, Fecteau and Jocelyn (2006) and Jocelyn (2006) tested balanced sampling to select a sample of businesses. Canadian unincorporated businesses complete their income tax returns either on paper or electronically. More than half of the returns are submitted electronically. Balanced sampling was used to select a sample from businesses that responded electronically so that, for some key variables that are known for the whole population, the sample means matched the known population means.

Balanced sampling can also be used to impute a missing value in case of item nonresponse. Indeed, using a model to predict an imputation allocates central values, which will lead to a biased inference on quantiles. In contrast, a random imputation generally increases the variances of the estimators. In order to solve this dilemma, Deville (1998, 2005, 2006) and Chauvet, Deville and Haziza (2010c, b) have proposed using imputation by prediction and to add a residual that is chosen amongst the residuals of the respondent according to a balanced sampling design. This is done to avoid adding a term of variance to the total of the imputed variable.

7.2 Balanced sampling versus other sampling techniques

Unequal probability sampling is a particular case of the cube method. Indeed, when the only auxiliary variable is the inclusion probability, the sample has a fixed sample size. The cube method is a generalization of the splitting method (see Deville and Tillé 1998), which includes several sampling algorithms with unequal probabilities (Brewer's method, pivotal method, corrected Sunter method, see Brewer 1975; Sunter 1977; Deville and Tillé 1998; Tillé 2006b). Stratification is also a particular case of balanced sampling. With the cube method, one can balance on overlapping strata and use qualitative and quantitative variables together. Systematic sampling can even be seen as a balanced sampling design on the order statistic related to the variable on which the population is ordered.

Almost all the other sampling techniques are particular cases of balanced sampling (except multistage sampling). In fact, balanced sampling is simply more general, in the sense that all the other methods of sampling can be implemented with the cube method. The cube method allows us to use any variable for balancing with a reasonable computation time. With the more general concept of balancing, strata can overlap, quantitative and qualitative variables can be used together, and the inclusion probabilities can be chosen freely.

It is well known that the ratio estimator and the post-stratified estimator are particular cases of the regression estimator. The regression estimator is also a particular case of the calibration estimator (which includes a non-linear adjustment). In the same way, balanced sampling is a more

general method of sampling that includes almost all the other methods. The algorithm of the cube method may seem complicated but, once implemented, it enables us to run a function with two arguments: the vector of inclusion probabilities and the matrix of balancing variables.

7.3 Choice of the balancing strategy

The main recommendation is to choose balancing variables that are closely correlated to the interest variables. As with any regression problem, the balancing variables must be chosen parsimoniously: one must not choose too many balancing variables because, accuracy no longer improves with a large number of variables and the instability of the variance estimator increases with each additional variable. Practically, the aim is not to estimate one variable but a set of interest variables. Thus, the set of auxiliary variables must be correlated to all the interest variables. Moreover, the auxiliary variables should not be too correlated amongst themselves.

Lesage (2008) has proposed a method to balance a sample on complex statistics rather than simply using population totals. The main idea consists in balancing on the linearized value (or influence function) of the parameter of interest. Breidt and Chauvet (2010b) have proposed using penalized balanced sampling in order to possibly relax some balancing constraints, which can be useful for instance in small domain estimation.

In many cases, the balancing variables contain measurement errors. For example, in most registers, one can suspect errors in the data. Missing values can obviously occur and auxiliary variables are often corrected by a method of imputation. As for calibration, the fact of having errors in the auxiliary variables is not very important as long as the calibration is done on the total of the auxiliary variables of the register. Indeed, with balanced sampling, the Horvitz-Thompson estimator is used and is unbiased even if the auxiliary variables are false. The gain in efficiency only depends on the correlation between the balancing variables and the interest variables. This correlation is rarely affected by errors in the balancing variables.

Several variables can be used to improve small domain estimates. To ensure that a domain D is not empty, one can simply add the auxiliary variable:

$$x_k = \begin{cases} \pi_k & \text{if } k \in D \\ 0 & \text{otherwise,} \end{cases}$$

which implies that the number of sampled units that belong to D is equal to

$$n_D = \sum_{k \in U} x_k = \sum_{k \in D} \pi_k,$$

if n_D is integer, or one of the closest two integers to n_D if n_D is not an integer.

In some cases, it is interesting to balance on auxiliary variables in subgroups, domains or strata. An interesting procedure described in Chauvet (2009) consists of separately running the flight phase in each stratum. A rounding problem will then occur in each stratum. These rounding problems can then be merged and a flight phase can be run again on the whole population. Finally, the landing phase is applied only to the whole population. This procedure enables us to roughly satisfy the balancing equations in each strata without cumulating the rounding problems.

The inclusion probabilities must be computed prior to sampling. When a linear model is assumed, these probabilities should in principle be proportional to the errors of the model in order to minimize variance (see Tillé and Favre 2005; Chauvet, Bonnery and Deville 2010a; Nedyalkova and Tillé 2009, 2010). This choice generalizes Neyman's allocation for stratified sampling (Neyman 1934). However, the inclusion probabilities often need to be chosen on others constraints. For instance, in order to construct the rotation groups of the French census, the inclusion probabilities must all be equal to a fifth.

7.4 Balancing versus calibration

Stratification is a particular case of balancing, while post-stratification is a particular case of calibration. In stratification and balancing, the weights do not become random. It is thus generally a better strategy. Nevertheless, more auxiliary information is needed for balancing. Indeed, for balanced sampling, the auxiliary variables must be known for all the units of the population, whereas, for calibration, only the population totals are needed. Balancing is a very interesting method for small population sizes. It is thus a very good method for selecting primary units in a multi-stage sampling design.

Both techniques can be used together. They are not contradictory. The best strategy consists of using balanced sampling and calibration together. Indeed calibration can resolve the small rounding problem that may remain after balancing. At the estimation stage, more auxiliary variables are often available because, in order to balance a sample, the auxiliary information must be known at the individual level while, in order to calibrate the sample, only the population totals are necessary.

Generally, it is recommended to re-calibrate on the balancing variables at the estimation stage even if more calibration variables are available. If only new variables are used in calibration, the effect of balancing can be lost. There is, however, one case where calibration can be used without re-calibrating on the balancing variables: when, conditionally on the calibration variables, we can reasonably assume that the balancing variables are no longer correlated to the variables of interest. This can occur when the balancing

and the calibration variables are the same variables measured at different moments, and the calibration variables are more recent.

When the determination coefficient between the interest variable and the auxiliary variables is equal to or close to one, then calibration is more efficient because of the rounding problem of balanced sampling. Anyway the most efficient strategy always consists of using balanced sampling and calibration together (see the simulation in Deville and Tillé 2004).

7.5 Accuracy of the balancing equations

It is possible to prove, under realistic assumptions (see Deville and Tillé 2004), that with the cube method

$$\left| \frac{\hat{X}_j - X_j}{X_j} \right| < O(p/n),$$

where p is the number of variables, and $O(x)/x$ is a quantity that remains bounded when x tends to infinity. With simple random sampling

$$\left| \frac{\hat{X}_j - X_j}{X_j} \right| = O_p(\sqrt{1/n}),$$

where $O_p(x)/x$ is a quantity that remains bounded in probability when x tends to infinity.

The gains in accuracy are therefore considerable. The small rounding problem can be fixed by a small calibration. The rounding problem comes from the fact that selecting a sample is an integer problem. It also occurs in stratification, which is a particular case of balancing. In stratification with proportional allocation, the sums of the inclusion probabilities in the strata are generally not integers. So, the sample sizes in the strata are obtained by rounding the sum of inclusion probabilities in the strata. The cube method does this rounding automatically and randomly in such a way as to ensure that the inclusion probabilities are exactly satisfied.

7.6 Balanced sampling in repeated surveys

An important difficulty occurs in repeated sampling. The problem comes from the fact that, when a balanced sample is selected with unequal inclusion probabilities, the complementary sample is not necessarily balanced. Indeed, the equality

$$\sum_{k \in S} \frac{x_k}{\pi_k} = \sum_{k \in U'} x_k$$

does not imply that

$$\sum_{k \in U \setminus S} \frac{x_k}{1 - \pi_k} = \sum_{k \in U} x_k.$$

This problem occurred in the French master sample. In this sampling design, the primary units, which are geographical

areas, are selected with unequal probabilities that are proportional to the size. After selecting the sample, some regions asked for complementary samples of areas that were not already selected. This question is intricate, because the complementary sample of a balanced sample is no longer balanced, and the aim is thus to select a balanced sample from a part of the population that is no longer balanced. Tillé and Favre (2004) gave a few methods to co-ordinate balanced samples, which were selected with unequal inclusion probabilities. More generally, the coordination (in the sense of managing overlap) of balanced samples can be difficult when the sampling design is balanced.

While challenging, it is possible to organize rotations if all the samples are selected together and the samples are selected with equal inclusion probabilities. Indeed, in this case the complementary $\bar{S} = U \setminus S$ of the samples S is also a balanced sample. A second balanced sample can be directly selected from \bar{S} and so on. This method was used to create five rotation groups in the French master sample. The five groups are five balanced samples of municipalities.

If the samples are selected with unequal inclusion probabilities, some solutions are described in Tillé and Favre (2004). An interesting particular case can easily be solved: when two non-overlapping samples must be selected with the same unequal inclusion probabilities $\pi_k < 0.5$ from the same population. First, a sample S_A balanced on \mathbf{x}_k must be selected with inclusion probabilities $\pi_{kA} = 2\pi_k$ such that

$$\sum_{k \in S_A} \frac{\mathbf{x}_k}{2\pi_k} = \sum_{k \in U} \mathbf{x}_k.$$

Next, a sample S_1 can be selected from S_A . This sample must be selected with inclusion probability $\pi_{kB} = 0.5$ and must be balanced on $\mathbf{x}_k/2\pi_k$, which gives the following balancing equations:

$$\sum_{k \in S_2} \frac{\mathbf{x}_k/(2\pi_k)}{1/2} = \sum_{k \in S_A} \frac{\mathbf{x}_k}{2\pi_k} = \sum_{k \in U} \mathbf{x}_k.$$

The sample $S_2 = S_A \setminus S_1$ is also balanced.

If the population changes over times (deaths and births), the organization of a rotation becomes much more difficult. This difficulty already occurs with stratified samples. Nevertheless, for stratification, several reasonable solutions exist (see, amongst others, De Ree 1999; Hesse 1998; Rivière 1999; Nedyalkova, Péa and Tillé 2006).

7.7 Main implementations of balanced sampling

An SAS/IML[®] implementation was first programmed by three students of the École nationale de la statistique et de l'analyse de l'information (Ensaï) (Bousabaa *et al.* 1999). An official version of the *Institut National de la Statistique et des Études Économiques* done by Tardieu (2001) and Rousseau and Tardieu (2004) is now available on the Insee Web site. Another SAS/IML[®] version done by Chauvet and

Tillé (2005b, a, 2006) is also available on the University of Neuchâtel Web site. In R language, the sampling package (Tillé and Matei 2007) allows us to use the cube method. These software programs are free, available over the Internet and are easy to use.

The available programs written using R language or SAS/IML[®] have no limit as far as population size is concerned. An application with 40 balanced variables is possible. In order to select the sample, the computation times increase with $N \times p^2$, where N is the population size and p the number of balancing variables. It is thus possible to select a sample in a population of several million statistical units.

Acknowledgements

This paper has been written in response to an invitation to speak at the Demographic Statistical Methods Division Seminar of the U.S. Census Bureau in June 2008. The author would like to thank the U.S. Census Bureau and particularly Patrick Flanagan without whom this paper would never have been written. The author is also grateful to an associate editor and two anonymous reviewers for valuable comments and corrections that helped to improve this paper.

References

- Ardilley, P. (1991). Échantillonnage représentatif optimum à probabilités inégales. *Annales d'Économie et de Statistique*, 23, 91-113.
- Ardilley, P. (2006). *Les Techniques de Sondage*. Technip, Paris.
- Bardaji, J. (2001). Un an après la sortie d'un contrat emploi consolidé : près de six chances sur dix d'avoir un emploi. *Premières Informations Synthèses, Direction de l'Animation de la Recherche des Études et des Statistiques (DARES) du Ministère du Travail des relations sociales et de la solidarité*, 43, 3, 1-8.
- Bertrand, P., Christian, B., Chauvet, G. and Grosbras, J.-M. (2004). Plans de sondage pour le recensement rénové de la population. In *Séries Insee Méthodes : Actes des Journées de Méthodologie Statistique*, Paris. Insee.
- Biggeri, L., and Falorsi, P.D. (2006). A probability sample strategy for improving the quality of the consumer price index survey using the information of the business register. In *Proceedings of the Conference of European Statisticians Group of Experts on Consumer Price Indices*, Eighth Meeting, Geneva, 10-12 May 2006.
- Bousabaa, A., Lieber, J. and Sirolli, R. (1999). La macro cube. Technical report, Ensaï, Rennes.
- Breidt, F.J., and Chauvet, G. (2010a). Improved variance estimation for balanced samples drawn via the cube method. *Journal of Statistical Planning and Inference*, 141, 479-487.
- Breidt, F.J., and Chauvet, G. (2010b). Penalized balanced sampling. Working paper, Ensaï.
- Brewer, K.R.W. (1975). A simple procedure for π pswor. *Australian Journal of Statistics*, 17, 166-172.
- Brewer, K.R.W. (1999). Design-based or prediction-based inference? Stratified random vs stratified balanced sampling. *International Statistical Review*, 67, 35-47.
- Chauvet, G. (2007). *Méthodes de Bootstrap en Population Finie*. PhD thesis, Université Rennes 2.
- Chauvet, G. (2009). Stratified balanced sampling. *Survey Methodology*, 35, 115-119.

- Chauvet, G., Bonnery, D. and Deville, J.-C. (2010a). Optimal inclusion probabilities for balanced sampling. *Journal of Statistical Planning and Inference*, 141, 2, 984-994.
- Chauvet, G., Deville, J. and Haziza, D. (2010b). Adapting the cube algorithm for balanced random imputation in surveys. Technical report, Ensai, Rennes.
- Chauvet, G., Deville, J. and Haziza, D. (2011). On balanced random imputation in surveys. *Biometrika*.
- Chauvet, G., and Tillé, Y. (2005a). *Fast SAS Macros for balancing Samples: user's guide*. Software Manual, University of Neuchâtel, <http://www2.unine.ch/statistics/page10890.html>.
- Chauvet, G., and Tillé, Y. (2005b). New SAS macros for balanced sampling. In *Journées de Méthodologie Statistique*, Insee, Paris.
- Chauvet, G., and Tillé, Y. (2006). A fast algorithm of balanced sampling. *Journal of Computational Statistics*, 21, 9-31.
- Chipperfield, J. (2009). An evaluation of cube sampling for ABS household surveys. Technical report, Australian Bureau of Statistics.
- Christine, M. (2006). Use of balanced sampling in the framework of the master sample for french household surveys. In *Joint Statistical Meeting of the American Statistical Association*, Seattle August 2006.
- Christine, M., and Wilms, L. (2003). Theoretical and practical problems related to the development of "EMEX": How to improve the precision of the regional supplements of National Surveys with an Additional Sample? In *Proceedings: Symposium 2003, Challenges in Survey Taking for the Next Decade*, Statistics Canada, Ottawa.
- Cumberland, W.G., and Royall, R.M. (1981). Prediction models in unequal probability sampling. *Journal of the Royal Statistical Society*, B, 43, 353-367.
- Cumberland, W.G., and Royall, R.M. (1988). Does simple random sampling provide adequate balance? *Journal of the Royal Statistical Society*, B, 50, 118-124.
- da Silva, A.D., da Silva Borges, A., Aires Leme, R. and Moura Reis Miceli, A.P. (2006). Modalidades alternativas de censo demográfico: o cenário internacional a partir das experiências dos estados unidos, França, Holanda, Israel e Alemanha. Technical report, Instituto Brasileiro de Geografia e Estatística.
- D'Alò, M., Di Consiglio, L., Falorsi, S. and Solari, F. (2006). Small area estimation of the Italian poverty rate. *Statistics in Transition*, 7, 771-784.
- De Ree, S.J.M. (1999). Co-ordination of business samples using measured response burden. In *Contributed paper, 52th Session of the ISI Helsinki*.
- Desplanques, G. (2000). La rénovation du recensement de la population. In *Actes de la séance du 5 octobre 2000 du séminaire méthodologique SFDS-Insee sur la rénovation du recensement*, 2-5.
- Dessertaine, A. (2006). Sondages et séries temporelles: une application pour la prévision de la consommation électrique. In *Actes des journées Françaises de Statistique 2006*, Clamart, France.
- Dessertaine, A. (2007). Sampling and data-stream: Some ideas to build balanced sampling using auxiliary Hilbertian informations. In *Proceedings of 56th the International Statistical Institute Conference: IPMS6 - New methods of sampling*, Lisboa, Portugal.
- Deville, J.-C. (1992). Constrained samples, conditional inference, weighting: Three aspects of the utilisation of auxiliary information. In *Proceedings of the Workshop on the Uses of Auxiliary Information in Surveys*, Örebro (Sweden).
- Deville, J.-C. (1998). La correction de la non-réponse par calage ou par échantillonnage équilibré. In *Recueil de la Section des méthodes d'enquêtes des communications présentées au 26^{ème} congrès de la Société Statistique du Canada*, 103-110, Sherbrooke.
- Deville, J.-C. (1999). Variance estimation for complex statistics and estimators: Linearization and residual techniques. *Survey Methodology*, 25, 193-203.
- Deville, J.-C. (2005). Imputation stochastique et échantillonnage équilibré. Technical report, École Nationale de la Statistique et de l'Analyse de l'Information.
- Deville, J.-C. (2006). Stochastic imputation using balanced sampling. In *Joint Statistical Meeting of the American Statistical Association*, Seattle August 2006.
- Deville, J.-C., Grosbras, J.-M. and Roth, N. (1988). Efficient sampling algorithms and balanced sample. In *COMPSTAT, Proceedings in Computational Statistics*, Heidelberg. Physica Verlag, 255-266.
- Deville, J.-C., and Tillé, Y. (1998). Unequal probability sampling without replacement through a splitting method. *Biometrika*, 85, 89-101.
- Deville, J.-C., and Tillé, Y. (2004). Efficient balanced sampling: The cube method. *Biometrika*, 91, 893-912.
- Deville, J.-C., and Tillé, Y. (2005). Variance approximation under balanced sampling. *Journal of Statistical Planning and Inference*, 128, 569-591.
- Dudoignon, L., and Vanheuverzwyn, A. (2006). Tirage d'un échantillon à probabilités inégales: application au panel Médiamat. In *Actes de des Journées de Méthodologie Statistique*, 1-10.
- Dumais, J., Bertrand, P. and Kauffmann, B. (2000). Sondage, estimation et précision dans la rénovation du recensement de la population. In *Actes de la séance du 5 octobre 2000 du séminaire méthodologique SFDS-Insee sur la rénovation du recensement*, 6-26.
- Dumais, J., and Isnard, M. (2000). Le sondage de logements dans les grandes communes dans le cadre du recensement rénové de la population. In *Séries Insee Méthodes: Actes des Journées de Méthodologie Statistique*, Paris. Insee, 100, 37-76.
- Durr, J.-M., and Dumais, J. (2001). Redesign of the french census of population. In *Proceedings: Symposium 2001, Achieving Data Quality in a Statistical Agency: A Methodological Perspective*, Statistics Canada, Ottawa.
- Durr, J.-M., and Dumais, J. (2002). Redesign of the french census of population. *Survey Methodology*, 28, 43-49.
- Even, K. (2002). Improved tool for evaluating employment and vocational training policy: Panel of beneficiaries. *Premières Informations Synthèses, Direction de l'Animation de la Recherche des Études et des Statistiques (DARES) du Ministère du Travail des relations sociales et de la solidarité*, 33, 1, 1-7.
- Falorsi, P.D., and Righi, P. (2008). A balanced sampling approach for multi-way stratification designs for small area estimation. *Survey Methodology*, 34, 223-234.
- Fecteau, S., and Jocelyn, W. (2006). Une application de l'échantillonnage équilibré: le plan de sondage des entreprises non incorporées. In *Méthodes d'enquêtes et sondages: pratiques européenne et nord-américaine*, (Eds., P. Lavallée and L.-P. Rivest), Paris. Dunod, 405-410.
- Fuller, W.A. (2009). Some design properties of a rejective sampling procedure. *Biometrika*, 96, 933-944.
- Fuller, W.A. (2010). Replication variance estimation for rejective sampling. In *Seminar of Statistics Canada*, June 2010, Ottawa.
- Gini, C. (1928). Une application de la méthode représentative aux matériaux du dernier recensement de la population italienne (1^{er} décembre 1921). *Bulletin of the International Statistical Institute*, 23, 2, 198-215.
- Gini, C., and Galvani, L. (1929). Di una applicazione del metodo rappresentativo all'ultimo censimento Italiano della popolazione (1^o dicembre, 1921). *Annali di Statistica*, Series 6, 4, 1-107.
- Gismondi, R. (2007). Quick estimation of tourist nights spent in Italy. *Statistical Methods and Applications*, 16, 141-168.

- Hájek, J. (1981). *Sampling from a Finite Population*. New York: Marcel Dekker.
- Hedayat, A.S., and Majumdar, D. (1995). Generating desirable sampling plans by the technique of trade-off in experimental design. *Journal of Statistical Planning and Inference*, 44, 237-247.
- Hesse, C. (1998). Sampling co-ordination: A review by country. Technical Report E9908, Direction des Statistique d'Entreprises, Insee, Paris.
- Jocelyn, W. (2006). Sampling and estimation strategies for the canadian unincorporated business population. In *Joint Statistical Meeting of the American Statistical Association*, Seattle August 2006.
- Kiaer, A. (1896). Observations et expériences concernant des dénombrements représentatifs. *Bulletin de l'Institut International de Statistique*, 9, 2, 176-183.
- Kiaer, A. (1899). Sur les méthodes représentatives ou typologiques appliquées à la statistique. *Bulletin de l'Institut International de Statistique*, 11, 1, 180-185.
- Kiaer, A. (1903). Sur les méthodes représentatives ou typologiques appliquées à la statistique. *Bulletin de l'Institut International de Statistique*, 13, 1, 66-78.
- Kiaer, A. (1905). Discours sans intitulé sur la méthode représentative. *Bulletin de l'Institut International de Statistique*, 14, 1, 119-134.
- Kott, P.S. (1986). When a mean-of-ratios is the best linear unbiased estimator under a model. *The American Statistician*, 40, 202-204.
- Langel, M., and Tillé, Y. (2010). Corrado Gini, a pioneer in balanced sampling and inequality theory. Technical report, University of Neuchâtel.
- Legg, J.C., and Yu, C.L. (2010). A comparison of sample set restriction procedures. *Survey Methodology*, 36, 69-79.
- Lesage, E. (2008). Contraintes d'équilibrage non linéaires. In *Méthodes d'enquêtes : applications aux enquêtes longitudinales, à la santé et aux enquêtes électorales*, (Eds., P. Guilbert, D. Haziza, A. Ruiz-Gazen and Y. Tillé), Paris. Dunod, 285-289.
- Marí, G., Barbará, G., Mitás, G. and Passamonti, S. (2007a). Construcción de un estimador de variancia para muestras balanceadas estratificadas. In *XXXV Coloquio Argentino de Estadística. Mar del Plata, Argentina. 22, 23 y 24 de Octubre de 2007*.
- Marí, G., Barbará, G., Mitás, G. and Passamonti, S. (2007b). Muestras equilibradas en poblaciones finitas: un estudio comparativo en muestras de explotaciones agropecuarias. In *Undécimas Jornadas "Investigaciones en la Facultad" de Ciencias Económicas y Estadística. noviembre de 2007*, Universidad Nacional de Rosario, Argentina.
- Nedyalkova, D., Péa, J. and Tillé, Y. (2006). A review of some current methods of coordination of stratified samples. introduction and comparison of new methods based on microstrata. Technical report, Université de Neuchâtel.
- Nedyalkova, D., and Tillé, Y. (2009). Optimal sampling and estimation strategies under linear model. *Biometrika*, 95, 521-537.
- Nedyalkova, D., and Tillé, Y. (2010). Bias robustness and efficiency in model-based inference. Technical report, University of Neuchâtel.
- Neyman, J. (1934). On the two different aspects of representative method: The method of stratified sampling and the method of purposive selection. *Journal of the Royal Statistical Society*, 97, 558-606.
- Neyman, J. (1952). *Lectures and Conferences on Mathematical Statistics and Probability*. Graduate School; U.S. Department of Agriculture, Washington.
- Périé, P. (2008). Échantillonnage à entropie maximale sous contraintes : un algorithme rapide basé sur l'optimisation linéaire en nombres binaires. In *Méthodes d'enquêtes : applications aux enquêtes longitudinales, à la santé et aux enquêtes électorales*, (Eds., P. Guilbert, D. Haziza, A. Ruiz-Gazen and Y. Tillé), Paris. Dunod, 294-299.
- Rivière, P. (1999). Coordination of samples: The microstrata methodology. In *13th International Roundtable on Business Survey Frames*, Paris. Insee.
- Rousseau, S., and Tardieu, F. (2004). La macro SAS CUBE d'échantillonnage équilibré, Documentation de l'utilisateur. Technical report, Insee, Paris.
- Royall, R.M. (1976a). Likelihood functions in finite population sampling theory. *Biometrika*, 63, 605-614.
- Royall, R.M. (1976b). The linear least squares prediction approach to two-stage sampling. *Journal of the American Statistical Association*, 71, 657-664.
- Royall, R.M. (1988). The prediction approach to sampling theory. In *Handbook of Statistics Volume 6: Sampling*, (Eds., P.R. Krishnaiah and C.R. Rao), Amsterdam. Elsevier/North-Holland, 399-413.
- Royall, R.M., and Pfeffermann, D. (1982). Balanced samples and robust bayesian inference in finite population sampling. *Biometrika*, 69, 401-409.
- Sunter, A. (1977). List sequential sampling with equal or unequal probabilities without replacement. *Applied Statistics*, 26, 261-268.
- Tardieu, F. (2001). Échantillonnage équilibré: de la théorie à la pratique. Technical report, Insee, Paris.
- Thionet, P. (1953). *La théorie des sondages*. Insee, Imprimerie nationale, Paris.
- Tillé, Y. (2001). *Théorie des sondages : échantillonnage et estimation en populations finies*. Dunod, Paris.
- Tillé, Y. (2006a). Balanced sampling by means of the cube method. In *Joint Statistical Meeting of the American Statistical Association*, Seattle August 2006.
- Tillé, Y. (2006b). *Sampling Algorithms*. New York: Springer.
- Tillé, Y., and Favre, A.-C. (2004). Co-ordination, combination and extension of optimal balanced samples. *Biometrika*, 91, 913-927.
- Tillé, Y., and Favre, A.-C. (2005). Optimal allocation in balanced sampling. *Statistics and Probability Letters*, 74, 31-37.
- Tillé, Y., and Matei, A. (2007). *The R Package Sampling*. The Comprehensive R Archive Network, <http://cran.r-project.org/>, Manual of the Contributed Packages.
- Tirari, M. (2006). Le plan de sondage équilibré et l'estimation du total d'une population finie. In *Méthodes d'enquêtes et sondages : pratiques européenne et nord-américaine*, (Eds., P. Lavallée and L.-P. Rivest), Paris, Dunod, 411-416.
- Valliant, R., Dorfman, A.H. and Royall, R.M. (2000). *Finite Population Sampling and Inference: A Prediction Approach*. New York: John Wiley & Sons, Inc.
- Wilms, L. (2000). Présentation de l'échantillon-maitre en 1999 et application au tirage des unités primaires par la macro cube. In *Séries Insee Méthodes : Actes des Journées de Méthodologie Statistique*, Paris. Insee.
- Yates, F. (1946). A review of recent statistical developments in sampling and sampling surveys. *Journal of the Royal Statistical Society*, A109, 12-43.
- Yates, F. (1960). *Sampling Methods for Censuses and Surveys*. Charles Griffin, London, England, third edition.

Innovations in survey sampling design: Discussion of three contributions presented at the U.S. Census Bureau

Jean Opsomer¹

1. Introduction

The U.S. Census Bureau is one of the largest survey data collection organizations in the world, in addition to its role in the collection of the U.S. Decennial Census data. The two major statistical tools used by the Census Bureau in designing its surveys are stratification and multi-stage sampling. These tools have been successfully implemented starting in the 1940s and have continually been adapted and refined since then.

While this general sampling approach has been very successful, there are increasing concerns about rising survey costs, decreasing response rates and new frame coverage issues (especially related to telephones). At the same time, advances in data collection methods, new data sources and computational tool offer opportunities for considering survey design approaches that would have been unfeasible before. In conjunction with the 2010 Redesign Program currently on-going at the Census Bureau, input was therefore sought from leading academic researchers in innovative sampling methods, as a way to initiate the exploration of possible new approaches to design surveys conducted by the Census Bureau. As a result, Profs. Steve Thompson (Simon Fraser University), Sharon Lohr (Arizona State University) and Yves Tillé (Université de Neuchâtel) were invited to give overview lectures on some of the designs they developed. I was invited to contribute a discussion to each of these lectures.

In the three sections that follow, I will summarize my comments to each of these lectures. My goals in those comments were to highlight the most important aspects of the sampling methods that were presented, to discuss some of the main opportunities for using these designs in the household sampling context, and to identify possible challenges in implementation.

2. Adaptive network and spatial sampling

Prof. Thompson's lecture covered a broad class of designs that includes adaptive cluster sampling, network sampling and adaptive web sampling. Unless I am referring to a specific design within this class, I will refer to these designs as "adaptive sampling" in what follows. A major

advantage of adaptive sampling is that it incorporates some of the features of "convenience" sampling approaches such as snowball sampling, including decreased reliance on a sampling frame and the ability to target sampling to portions of the population of particular interest. But unlike convenience sampling, adaptive sampling remains firmly design-based, in the sense of allowing randomization-based finite population estimation and inference.

In adaptive sampling procedures, an initial sample s_0 is drawn according to a probability sampling design $p_0(s_0)$. Based on the characteristics of the elements in s_0 (e.g., presence/absence of features of interest or an enumeration of "links" to other elements in the population), a follow-up sample s_1 is selected from the remaining population, using a conditional sampling design $p_1(s_1 | s_0)$. This process is repeated with successive incremental samples s_2, s_3, \dots until a target criterion such as overall sample size or number of sampling "waves" is reached, and the final sample is the union of each of the successive samples. The specifics on how the waves are drawn varies by adaptive design. Section 2.2 of Thompson's article in this issue and Thompson (2006) provide additional details for adaptive web sampling, a very flexible type of adaptive sampling that includes many of the other designs as special cases.

Because the designs for each of the sampling waves are probability designs, it is possible to obtain valid design-based estimators. A simple estimator for the finite population mean $\mu_N = N^{-1} \sum_U y_i$ is constructed as follows. Based on the initial design p_0 with associated inclusion probabilities π_{0i} , an unbiased estimator for the population mean is given by $\hat{\mu}_0 = N^{-1} \sum_{s_0} y_i / \pi_{0i}$. For each of the subsequent waves $k = 1, \dots, K$, an unbiased estimator of μ_N is given by $z_k = \sum_{s_{k-1}} y_i + \sum_{s_k} y_i / q_{ki}$, where q_{ki} are conditional inclusion probabilities for wave k (see Thompson (2006) for details on construction of the q_{ki} , and Section 2.4 of Thompson's article in this issue for specific examples). Letting $\hat{\mu}_r = K^{-1} \sum_{k=1}^K z_k$, an unbiased estimator for μ_N is obtained as $\hat{\mu} = w \hat{\mu}_0 + (1 - w) \hat{\mu}_r$, which is a linear combination of the initial estimator and the mean of the subsequent estimators.

The estimator $\hat{\mu}$ is design unbiased but it depends on the order of the waves in which the sample was obtained. A more precise estimator can be obtained by averaging over all the different orders in which the same sample could have

1. Jean Opsomer, Department of Statistics, Colorado State University, Fort Collins, CO 80523-1877. E-mail: jopsomer@stat.colostate.edu.

been obtained. For small sample sizes, an explicit expression is available for this more efficient estimator, but in general it needs to be approximated by repeated sampling from an appropriately defined Markov chain, and taking the mean of the samples. The exact methods for setting up the chain and drawing the samples are described in Thompson (2006), which also discusses variance estimation for the resulting estimator.

One of the primary advantages of adaptive sampling designs is that they allow the survey organization to focus the sample in portions of interest in the population. This is particularly useful in situations where some of the elements of interest are relatively rare and where they cannot be identified *a priori* in a sampling frame. Examples of such situations are surveys of hunting and fishing behavior, recent immigrants, home-schoolers, or owners of family-owned businesses. In each of these cases, the elements are quite “diffuse” in the population and no comprehensive frame is generally available. However, it is likely that individuals who are part of this population will be able to provide information on other individuals, so that links can be identified and sampled across different adaptive sampling waves. Note that adaptive sampling can also be used when these types of rare elements are part of a subpopulation of interest within a survey of a larger and non-rare population. For instance, a survey of school children might want to include a stratum of home-schooled children.

Finding relatively rare (sub)populations is a common challenge in surveys, and a number of methods are regularly deployed to deal with this issue. Perhaps the most common sampling design in the context of household surveys is stratified multi-stage sampling. To the extent that relevant PSU-level auxiliary information is available, the survey organization can oversample PSU expected to contain a larger fraction of the groups of interest. An example of such a situation is a survey of African-American males at risk of Parkinson’s disease, in which Census tracts with higher African-American population fraction could be oversampled. Another sampling design that can be useful in this context is multi-phase sampling. In this case, the first phase of sampling is used either as a screening sample or as a way to collect relevant auxiliary information, while subsequent phases focus on obtaining the survey data of interest. The Agricultural Resource Management Survey (ARMS) conducted by the USDA follows this design. A sample of all farms is selected in phase 1, in which farm characteristics for the survey year are collected. In later phases, targeted sampled based on the commodities of interest (*e.g.*, dairy, wheat, *etc*) are selected. A third sampling approach that is sometimes useful for obtaining samples of rare (sub)populations is multi-frame sampling. The principle underlying multi-frame sampling is to combine several frames with

different coverage characteristics, for instance a “good” frame containing a large proportion of elements of interest but potentially incomplete and a “bad” frame that is comprehensive but contains a low proportion of elements of interest. For instance, a survey of companies in a particular industry might be able to use an industry group membership list as the “good” frame and a general company list as the “bad” frame. For a more in-depth look at multi-frame sampling, see Section 3 below.

Compared to these three designs, adaptive sampling is more flexible and allows finer control over the number and characteristics of elements that are included in the sample, which will often result in improved efficiency and/or lower cost. A drawback of adaptive sampling is that information needs to be collected on the linkages between elements, which can increase respondent burden and data collection cost, and potentially raises confidentiality issues.

Because adaptive sampling frequently relies on “links” between elements in order to define the conditional selection probabilities in the sampling waves, it is also particularly well-suited for surveys that are interested in studying connections between elements in a population. Examples of such situations might be surveys involving transactions or relationships between businesses, surveys of barter/trading behavior of households, and surveys of family network relationships or characteristics.

For a survey organization contemplating adoption of adaptive sampling, a number of issues related to estimation and data dissemination need to be considered. In many cases, the survey data are released in the form of a weighted dataset, and variance estimates are provided in the form of a simplified design description (*e.g.*, strata and PSUs), replicate weights or generalized variance functions. It is also very common for the weights to be calibrated and/or adjusted for non-response. Estimators for adaptive designs are indeed expressible as weighted sample sums, so that a weighted dataset could readily be created even for the Markov chain version of the estimators mentioned above. The choice of how to best provide variance estimates with the dataset is something that still needs to be investigated and might depend on the specifics of the survey. Similarly, how to incorporate calibration and nonresponse adjustments in adaptive sampling estimation is an area where additional research is needed.

3. Sampling with multiple overlapping frames

Prof. Lohr gave a comprehensive overview of general sampling designs and estimation methods when sampling uses multiple frames. Traditional approaches for conducting surveys are increasingly called into question today, because

of increasing costs, decreasing response levels for traditional modes, and increasing concerns for undercoverage of existing sampling frames (e.g., landline telephone numbers reached by RDD). By drawing samples from several frames instead of from a single frame, it is possible to reduce survey costs, improve the coverage of the overall sample, and potentially even increase response rates depending on the specific survey being conducted (for instance, because of improved respondent identifier information in one of the frames).

Multiple frame sampling is a pure randomization-based approach to draw samples, and sampling within the individual frames follows the same methodology as “classical” single-frame sampling. Fully design-based estimation methods for multiple-frame sampling are available, several of which can readily be deployed in the large-scale survey context in which a weighted dataset is the primary output (see below). The key feature of all estimation methods is the estimation of the frame overlap, which is typically unknown but needs to be accounted for. This is done by, for each frame, constructing design-based estimators for the subpopulation(s) of elements that also fall in the other frame(s). The estimators for the characteristics of the frame intersection(s) then need to be combined across frames. Existing methods differ in how they combine these estimators, with the simplest methods using sample-size weighted averages and more complex estimators weighting by estimates of the precision of the individual estimators.

Sampling from multiple frames is particularly applicable in cases where no single frame is available that covers the whole population. Typical examples of such situations are RDD sampling, where an increasing fraction of the population is not reachable through a landline telephone number, surveys of professionals or businesses with partial listings available from vendors or professional organizations. Other situations in which multiple frame sampling might be applicable are surveys of rare subpopulations that exist within a larger population. An overall frame for the population exists, but screening respondents for whether they belong to the subpopulation is time-consuming and expensive. An alternate frame containing a much higher proportion of elements from the subpopulation of interest is sometimes available, but if the coverage of that frame is incomplete, the survey organization might not be willing to rely on it for fear of not obtaining a valid sample. Combining the alternate but incomplete subpopulation frame with the complete but inefficient population frame might be both cost-effective and statistically defensible. Examples of surveys of such subpopulations are surveys of hunting and fishing, where a license frame often exists but it might be incomplete or out of date. This multiple frame approach might also be useful for a survey of the general population,

as a way to increase the sample size within certain subpopulations of particular interest. For instance, in a general survey of farms, it might be of interest to produce estimates for organic farms, which only represent a small fraction of farms but with many of those listed in organic business directories. Section 1 of Lohr’s article in this issue gives several additional examples of the wide applicability of multiple frame surveys.

As noted above, estimation methods involve the construction of estimators for the frame intersection subpopulation, which requires selection of a weighting method for the estimators obtained from the different frames. Weighting methods that rely on estimating the precision of these estimators might be preferred from an efficiency perspective. However, they are somewhat problematic to implement in practice, because the resulting weights can vary for different variables in the survey. More practical approaches will forego some efficiency in order to be able to have single weights for all survey variables, a key feature emphasized repeatedly in Lohr’s article in this issue. The *pseudo-maximum likelihood* (PML) method of Skinner and Rao (1996) produces a single set of weights and is recommended by Lohr as the method of choice for single surveys, while a simpler fixed-weight approach is preferable for longitudinal surveys.

While the basic methodology for constructing design-based estimators for multiple frame sampling is in place today, there is still a need for further research in approaches for applying calibration and nonresponse adjustment in this context. Because it is possible to apply those adjustments at the individual frame level, the population level, or both levels (depending on the available auxiliary information), an investigation of the properties of the estimators under these different scenarios would be very useful, and should be used to develop guidelines for survey practitioners. Section 3 of Lohr’s article in this issue discusses some initial results in this area.

Variance estimation methods for multiple-frame estimators have been developed and are reviewed in Section 4.2 of Lohr’s article, and include both linearization and replication approaches. An important practical issue in the use of the linearization approach is that it requires access to the frame identification for all the elements in the sample, because it involves separate estimation of the variance in each frame. This might be undesirable for the survey organization producing the data, for reasons of data confidentiality. In the case of replication methods such as jackknife and bootstrap, it is possible for the survey organization to create sets of replicate weights that do not require disclosure of the frame identity of individual sample elements to the data users. Lohr (2007) recommends the *combined bootstrap* approach for inference for multiple frame sampling.

As an alternative, the *grouped jackknife* of Kott (2001) could also be considered.

Implementing multiple frame sampling surveys can be more challenging than single-frame surveys. There needs to be awareness for the increased potential for nonsampling errors, as discussed in Section 5 of Lohr's article, especially if the data collection modes or protocols vary across frames. For instance, sampled elements in one frame get an advance letter, while those in another frame receive a "cold call" because of lack of address information. It is also possible that the nonresponse characteristics differ across frame, so that separate adjustments are required. Finally, in many cases the elements present in the different frames might have different characteristics (e.g., organic farms belonging to a national organic business association vs. those that do not). In all those cases, attention to frame-specific effects and careful weight construction are required in order to obtain valid survey estimators. On the other hand, the presence of multiple frames provides opportunities for measuring nonsampling errors, because they entail multiple samples from the same population. For instance, it might be useful to perform "cold calls" for a portion of the selected elements in the frame with addresses to evaluate mode effects.

4. Balanced sampling with the cube method

The presentation by Prof. Tillé covered the fundamentals of balanced sampling and described the *cube method*, which he developed as a practical algorithm implementing the drawing of balanced samples. The goals of balanced sampling designs are to maintain the representation of the population structure in the sample (hence the term "balance"), and to improve the efficiency of survey estimators. Today, most survey statisticians apply stratification as the primary tool to achieve these two goals. Stratification achieves balance by forcing the sample composition to match the stratum allocation, and improves the efficiency of estimators by removing the component of variance due to between-stratum differences. Systematic sampling is also used to achieve these goals, most commonly in natural resource surveys. In this case, the sample composition matches the population composition exactly along the sorting variable, and approximately for any variable correlated with the sorting variable. Efficiency is gained because sample moments of the variables of interest (approximately) match population moments. While both approaches are widely used and work well, they are relatively inflexible. Stratification often involves dividing the population into "cells" defined by the intersection of stratification variables, which might lead to a proliferation of many small cells with

corresponding small sample sizes. Systematic sampling is a highly constrained form of sampling with limited amount of flexibility in sample construction, and with the additional issue of the lack of a design-based variance estimator.

Balanced sampling can be viewed as a generalization of stratification. Under this interpretation, stratified samples are drawn with given probabilities of inclusion for all the population elements, but subject to constraints on the sample size in each stratum. In balanced sampling, the stratification constraints are replaced by constraints of the form $\sum_s x_i / \pi_i = \sum_U x_i$, where x_i is a vector of *balancing variables*. When the x_i are stratum indicators, balanced sampling is the same as stratification, but any categorical or continuous variables (or combination thereof) can be used, which provides a high degree of flexibility in sample construction.

As noted above, the cube method is an algorithm that draws balanced samples given a set of inclusion probabilities and constraints. If exactly balanced samples exist in the population, the algorithm will try to select one of them. If no sample can be found that has the postulated inclusion probabilities and satisfies the balancing constraints exactly, it will attempt to come as close as possible to satisfying the constraints. The cube method requires that the balancing variables x_i be known for all elements in the population. Depending on the survey context, this requirement might represent a key limitation on the applicability of balanced sampling.

Despite the fact that balancing on population-level auxiliary variables is done at the design stage, it seems likely that in practice, calibration and other weight adjustments such as for nonresponse will still often be required. In fact, Tillé recommends the combination of balancing and calibration as the most efficient strategy (see Section 7.4 of Tillé's article in this issue). The theoretical properties of estimators that are both balanced and calibrated still needs to be fully worked out, however.

While balanced sampling maintains the inclusion probabilities of the elements in the population, it is clear that the presence of the balancing constraints affects their *joint* inclusion probabilities and hence the variance of the estimators. This topic is addressed in Section 6 of Tillé's article. Deville and Tillé (2005) showed that, under certain conditions, the variance of balanced sampling estimators can be approximated by a linearization-type variance, which depends on the residuals of a linear regression of the survey variables on the balancing variables. While this is an important and useful result, it does not lead to a variance estimation approach that is applicable to all survey applications. One issue is that variance estimation based on this method requires access to the balancing variables for all the survey respondents, and these might not be made

publicly available as part of the survey dataset. In this context, a replication-based method might be particularly attractive, because it would not require releasing these variables. However, no such method is currently available.

Balanced sampling has close connections with *rejective sampling*, which aims to achieve the same goals. In rejective sampling, a sample is drawn with prespecified inclusion probabilities and the sample is accepted or rejected based on whether it is within a given tolerance level of a balancing constraint. If the sample is rejected, the procedure is repeated until a sample is found that falls within the tolerance level. While rejective sampling has a long history, Fuller (2009) described some asymptotic theory that showed that asymptotically, his version of rejective sampling was approximately equivalent to balanced sampling.

5. Closing remarks

The methods covered in the three lectures are remarkably complementary. Adaptive designs make it possible to obtain randomization-based, statistically valid samples for populations that have traditionally been difficult to sample efficiently. Very little frame information is required to draw such a sample, but a significant amount of effort has to be expended during the data collection in order to identify and follow the “links” among the elements, and draw the successive samples. In contrast, balanced sampling is useful when very detailed frame information is available, and in that situation, it allows for highly customized and efficient sample designs. Once a balanced sample is drawn, the data collection can proceed in the same manner as for traditional

surveys. Multiple frame sampling covers an intermediate case, in the sense that no single good frame exists but several partial frames are used to “offset” each other’s weaknesses. Separate samples are drawn from each frame, and data collection proceeds as usual, except for that fact that it is necessary to determine which frame(s) each sampled respondent belong to.

Combined with the existing approaches already in use, these three new sampling methods have the potential to greatly increase the flexibility with which samples can be customized for specific applications, to reduce survey costs and to increase the precision of survey estimators.

References

- Deville, J.-C., and Tillé, Y. (2005). Variance approximation under balanced sampling. *Journal of Statistical Planning and Inference*, 2, 569-591.
- Fuller, W.A. (2009). Some design properties of a rejective sampling procedure. *Biometrika*, 96(4), 933-944.
- Kott, P.S. (2001). The delete-a-group jackknife. *Journal of Official Statistics*, 17, 521-526.
- Lohr, S. (2007). Recent developments in multiple frame surveys. In *ASA Proceedings of the Joint Statistical Meetings*, American Statistical Association, 3257-3264.
- Skinner, C.J., and Rao, J.N.K. (1996). Estimation in dual frame surveys with complex designs. *Journal of the American Statistical Association*, 91, 349-356.
- Thompson, S.K. (2006). Adaptive web sampling. *Biometrics*, 62, 1224-1234.

ACKNOWLEDGEMENTS

Survey Methodology wishes to thank the following people who have provided help or served as referees for one or more papers during 2011.

- M. Aitkin, *University of Melbourne*
 J.-F. Beaumont, *Statistics Canada*
 W. Bell, *U.S. Census Bureau*
 E. Berg, *Iowa State University*
 Y. Berger, *University of Southampton*
 P. Biemer, *Research Triangle Institute*
 C. Bocci, *Statistics Canada*
 J. van den Brakel, *Statistics Netherlands*
 J.M. Brick, *Westat Inc*
 J. Bushery, *U.S. Census Bureau*
 P. Cantwell, *U.S. Bureau of the Census*
 R. Chambers, *Centre for Statistical and Survey Methodology*
 R. Creecy, *U.S. Census Bureau*
 M. Di Zio, *Italian National Statistical Institute*
 P. Dick, *Statistics Canada*
 Y. Dong, *Temple University*
 A.H. Dorfman, *U.S. Bureau of Labour Statistics*
 G. Dubreuil, *Statistics Canada*
 S. Eckman, *Institute for Employment Research*
 A. Eideh, *Al-Quds University*
 M. Elliott, *University of Michigan*
 J.L. Eltinge, *U.S. Bureau of Labor Statistics*
 V. Estevao, *Statistics Canada*
 G. Forsman, *Swedish Transport Administration*
 W.A. Fuller, *Iowa State University*
 J. Gambino, *Statistics Canada*
 S. Godbout, *Statistics Canada*
 H.A. Gutiérrez Rojas, *Faculty of Statistics – Universidad Santo Tomás*
 D. Haziza, *Université de Montréal*
 Y. He, *Harvard Medical School*
 S. Heeringa, *University of Michigan*
 M. Hidirolou, *Statistics Canada*
 B. Hulliger, *University of Applied Sciences Northwestern Switzerland*
 P.M. Joyce, *U.S. Bureau of Census*
 D. Judkins, *Westat Inc*
 D. Kasprzyk, *NORC at the University of Chicago*
 J.-K. Kim, *Iowa State University*
 P. Kott, *Research Triangle Institute*
 T. Krenzke, *Westat Inc.*
 P. Lahiri, *JPSM, University of Maryland*
 P. Lavallée, *Statistics Canada*
 H. Lee, *Westat Inc.*
 J. Legg, *Amgen Inc., U.S.A.*
 J. Li, *Westat, Inc.*
 S. Lohr, *Arizona State University*
 E. López Escobar, *University of Southampton*
 P. Lynn, *University of Essex*
 D.J. Malec, *National Center for Health Statistics*
 H. Mantel, *Statistics Canada*
 K. Miller, *U.S. National Center for Health Statistics*
 J.M. Montaquila, *Westat Inc.*
 F. Moura, *Universidade Federal de Rio de Janeiro*
 T. Mulcahy, *National Opinion Research Center*
 J.F. Muñoz Rosas, *University of Granada*
 B. Nandram, *WPI*
 G. Nathan, *Hebrew University*
 D. Nelson, *Center for Chronic Disease Outcomes Research*
 S. Oman, *Hebrew University*
 J. Opsomer, *Colorado State University*
 J.L. Parsons, *USDA, National Agricultural Statistics Service*
 Z. Patak, *Statistics Canada*
 D. Pfeffermann, *Hebrew University*
 C. Poirier, *Statistics Canada*
 N.G.N. Prasad, *University of Alberta*
 J. Preston, *ABS*
 S. Rabe-Hesketh, *UC Berkeley*
 J.N.K. Rao, *Carleton University*
 J. Reiter, *Duke University*
 L.-P. Rivest, *Université Laval*
 S. Rubin-Bleuer, *Statistics Canada*
 N. Schenker, *National Center for Health Statistics*
 F.J. Scheuren, *National Opinion Research Center*
 B. Schouten, *Statistics Netherlands*
 R. Sigman, *Westat, Inc.*
 P. do N. Silva, *Escola Nacional de Ciências Estatísticas*
 M. Sinclair, *NORC at the University of Chicago*
 A. Singh, *National Opinion Research Center*
 B. Skalland, *National Opinion Research Center*
 E. Sluid, *University of Maryland*
 P. Smith, *Office for National Statistics*
 E. Stasny, *Ohio State University*
 D. Steel, *University of Wollongong*
 L. Stokes, *Southern Methodist University*
 M. Sverchkov, *Bureau of Labor Statistics*
 M. Thompson, *University of Waterloo*
 C. Tucker, *Consultant*
 N. Tzavidis, *University of Southampton*
 B. van der Klaauw, *University of Amsterdam*
 J. van der Laan, *Statistics Netherlands*
 V.J. Verma, *Università degli Studi di Siena*
 J. Wang, *Hewlett-Packard Labs*
 Z. Wang, *Sir Wilfrid Laurier University*
 K.M. Wolter, *National Opinion Research Center*
 J. Wood, *Office for National Statistics*
 C. Wu, *University of Waterloo*
 Y. You, *Statistics Canada*
 C. Yu, *Iowa State University*
 W. Yung, *Statistics Canada*
 A. Zaslavsky, *Harvard Medical School*
 L.-C. Zhang, *Statistics Norway*

Acknowledgements are also due to those who assisted during the production of the 2011 issues: Céline Ethier of Statistical Research and Innovation Division, Christine Cousineau and Teresa Jewell of Household Survey Methods Division, Nick Budko and Sophie Chartier of Business Survey Methods Division, Dominique Lavoie of Social Survey Methods Division, Matthew Belyea, Louise Demers, Anne-Marie Fleury, Roberto Guido, Lydia Kokline, Liliane Lanoie, Darquise Pellerin, Joseph Prince and Fadi Salibi of Dissemination Division.

ANNOUNCEMENTS

Nominations Sought for the 2013 Waksberg Award

The journal *Survey Methodology* has established an annual invited paper series in honour of Joseph Waksberg to recognize his contributions to survey methodology. Each year a prominent survey statistician is chosen to write a paper that reviews the development and current state of an important topic in the field of survey methodology. The paper reflects the mixture of theory and practice that characterized Joseph Waksberg's work.

The recipient of the Waksberg Award will receive an honorarium from Westat. The paper will be published in a future issue of *Survey Methodology*.

The author of the 2012 Waksberg paper will be selected by a four-person committee appointed by *Survey Methodology* and the American Statistical Association. Nomination of individuals to be considered as authors or suggestions for topics should be sent before February 28, 2012 to the chair of the committee, Mary Thompson (methomps@uwaterloo.ca).

Previous Waksberg Award honorees and their invited papers are:

- 2001 Gad **Nathan**, "Telesurvey methodologies for household surveys – A review and some thoughts for the future?". *Survey Methodology*, vol. 27, 1, 7-31.
- 2002 Wayne A. **Fuller**, "Regression estimation for survey samples". *Survey Methodology*, vol. 28, 1, 5-23.
- 2003 David **Holt**, "Methodological issues in the development and use of statistical indicators for international comparisons". *Survey Methodology*, vol. 29, 1, 5-17.
- 2004 Norman M. **Bradburn**, "Understanding the question-answer process". *Survey Methodology*, vol. 30, 1, 5-15.
- 2005 J.N.K. **Rao**, "Interplay between sample survey theory and practice: An appraisal". *Survey Methodology*, vol. 31, 2, 117-138.
- 2006 Alastair **Scott**, "Population-based case control studies". *Survey Methodology*, vol. 32, 2, 123-132.
- 2007 Carl-Erik **Särndal**, "The calibration approach in survey theory and practice". *Survey Methodology*, vol. 33, 2, 99-119.
- 2008 Mary E. **Thompson**, "International surveys: Motives and methodologies". *Survey Methodology*, vol. 34, 2, 131-141.
- 2009 Graham **Kalton**, "Methods for oversampling rare subpopulations in social surveys". *Survey Methodology*, vol. 35, 2, 125-141.
- 2010 Ivan P. **Fellegi**, "The organisation of statistical methodology and methodological research in national statistical offices". *Survey Methodology*, vol. 36, 2, 123-130.
- 2011 Danny **Pfeffermann**, "Modelling of complex survey data: Why model? Why is it a problem? How can we approach it?". *Survey Methodology*, vol. 37, 2, 115-136.
- 2012 Lars **Lyberg**, Manuscript topic under consideration.

Members of the Waksberg Paper Selection Committee (2011-2012)

Mary Thompson, *University of Waterloo* (Chair)

J.N.K. Rao, *Carleton University*

Steve Heeringa, *University of Michigan*

Cynthia Clark, *USDA*

Past Chairs:

Graham Kalton (1999 - 2001)

Chris Skinner (2001 - 2002)

David A. Binder (2002 - 2003)

J. Michael Brick (2003 - 2004)

David R. Bellhouse (2004 - 2005)

Gordon Brackstone (2005 - 2006)

Sharon Lohr (2006 - 2007)

Robert Groves (2007 - 2008)

Leyla Mojadjer (2008 - 2009)

Daniel Kasprzyk (2009 - 2010)

Elizabeth A. Martin (2010 - 2011)

JOURNAL OF OFFICIAL STATISTICS

An International Review Published by Statistics Sweden

JOS is a scholarly quarterly that specializes in statistical methodology and applications. Survey methodology and other issues pertinent to the production of statistics at national offices and other statistical organizations are emphasized. All manuscripts are rigorously reviewed by independent referees and members of the Editorial Board.

Contents Volume 27, No. 2, 2011

Preface	
Annelies Blom, Frauke Kreuter.....	151
Proxy Pattern-Mixture Analysis for Survey Nonresponse	
Rebecca R. Andridge, Roderick J.A. Little.....	153
Imputation and Estimation under Nonignorable Nonresponse in Household Surveys with Missing Covariate Information	
Danny Pfeffermann, Anna Sikov	181
An Analysis of Nonignorable Nonresponse to Income in a Survey with a Rotating Panel Design	
Caterina Giusti, Roderick J.A. Little	211
Indicators for Monitoring and Improving Representativeness of Response	
Barry Schouten, Natalie Shlomo, Chris Skinner.....	231
A Pseudo-GEE Approach to Analyzing Longitudinal Surveys under Imputation for Missing Responses	
Iván A. Carrillo, Jiahua Chen, Changbao Wu.....	255
Effects of Increasing the Incentive Size in a Longitudinal Study	
Willard L. Rodgers	279
Attrition in the Swiss Household Panel: Is Change Associated with Drop-out?	
Marieke Voorpostel, Oliver Lipps.....	301
Keeping Track of Panel Members: An Experimental Test of a Between-Wave Contact Strategy	
Katherine A. McGonagle, Mick P. Couper, Robert F. Schoeni	319
Using Paradata and Other Auxiliary Data to Examine Mode Switch Nonresponse in a "Recruit-and-Switch" Telephone Survey	
Joseph W. Sakshaug, Frauke Kreuter.....	339
Interviewer Effects on Nonresponse in the European Social Survey	
Annelies G. Blom, Edith D. de Leeuw, Joop J. Hox	359
Toward a Benefit-Cost Theory of Survey Participation: Evidence, Further Tests, and Implications	
Eleanor Singer.....	379
Applying Motivation Theory to Achieve Increased Response Rates, Respondent Satisfaction and Data Quality	
Marika Wenemark, Andreas Persson, Helle Noorlind Brage, Tommy Svensson, Margareta Kristenson.....	393

All inquiries about submissions and subscriptions should be directed to journals@scb.se

JOURNAL OF OFFICIAL STATISTICS

An International Review Published by Statistics Sweden

JOS is a scholarly quarterly that specializes in statistical methodology and applications. Survey methodology and other issues pertinent to the production of statistics at national offices and other statistical organizations are emphasized. All manuscripts are rigorously reviewed by independent referees and members of the Editorial Board.

Contents Volume 27, No. 3, 2011

A Unit-Error Theory for Register-Based Household Statistics Li-Chun Zhang.....	415
Using Statistical Models for Sample Design of a Reinterview Program Jianzhu Li, J. Michael Brick, Bac Tran, Phyllis Singer	433
Centre Sampling Technique in Foreign Migration Surveys: A Methodological Note Gianluca Baio, Gian Carlo Blangiardo, Marta Blangiardo.....	451
Algorithms for Correcting Sign Errors and Rounding Errors in Business Survey Data Sander Scholtus.....	467
Bayesian Estimation of Population Size via Linkage of Multivariate Normal Data Sets Brunero Liseo, Andrea Tancredi	491
Random Group Variance Estimators for Survey Data with Random Hot Deck Imputation Jun Shao, Qi Tang.....	507
Statistical Properties of Multiplicative Noise Masking for Confidentiality Protection Tapan K. Nayak, Bimal Sinha, Laura Zayatz	527
Book Review	545
In Other Journals	551

All inquiries about submissions and subscriptions should be directed to jos@scb.se

CONTENTS

TABLE DES MATIÈRES

Volume 39, No. 2, June/juin 2011

Alison L. Gibbs, Kevin J. Keen and Liqun Wang Case studies in data analysis	181
Xiaoqing Niu, Pengfei Li and Peng Zhang Testing homogeneity in a multivariate mixture model.....	218
Aleksey Min and Claudia Czado Bayesian model selection for D-vine pair-copula constructions	239
Giampiero Marra and Rosalba Radice Estimation of a semiparametric recursive bivariate probit model in the presence of endogeneity	259
Richard Charnigo and Cidambi Srinivasan Self-consistent estimation of mean response functions and their derivatives	280
Seo Young Park and Yufeng Liu Robust penalized logistic regression with truncated loss functions	300
Huazhen Lin, Ling Zhou, Heng Peng and Xiao-Hua Zhou Selection and combination of biomarkers using ROC method for disease classification and prediction.....	324
Stefan H. Steiner, Nathaniel T. Stevens, Ryan Browne and R. Jock Mackay Planning and analysis of measurement reliability studies	344
Shojaeddin Chenouri, Christopher G. Small and Thomas J. Farrar Data depth-based nonparametric scale tests	356
Minqiang Li, Liang Peng and Yongcheng Qi Reduce computation in profile empirical likelihood method	370

Volume 39, No. 3, September/septembre 2011

Special Issue: Special Issue in Honour of Jack Kalbfleisch and Jerry Lawless

Richard J. Cook and Grace Y. Yi A special issue of CJS in honour of Jack Kalbfleisch and Jerry Lawless	385
Jiahua Chen and Pengfei Li Tuning the EM-test for finite mixture models	389
Niels Keiding Age-period-cohort analysis in the 1870s: Diagrams, stereograms, and the basic differential equation	405
Daeyoung Kim and Bruce G. Lindsay Modal simulation and visualization in finite mixture models	421
Terry C.K. Lee, Leilei Zeng, Darby J.S. Thompson and C.B. Dean Comparison of imputation methods for interval censored time-to-event data in joint modelling of tree growth and mortality	438
Ni Li, Do-Hwan Park, Jianguo Sun and KyungMann Kim Semiparametric transformation models for multivariate panel count data with dependent observation process	458
Karen McKeown and Nicholas P. Jewell Current status observation of a three-state counting process with application to simultaneous accurate and diluted HIV test data	475
John M. Neuhaus and Charles E. McCulloch The effect of misspecification of random effects distributions in clustered data settings with outcome-dependent sampling	488
Ross L. Prentice and Ying Huang Measurement error modeling and nutritional epidemiology association analyses	498
Jing Qin and Biao Zhang Optimal estimating functions in incomplete data and length biased sampling data problems	510
Alastair J. Scott and Chris J. Wild Fitting regression models with response-biased samples	519
Yang Yang and Vijayan N. Nair Parametric inference for time-to-failure in multi-state semi-Markov models: A comparison of marginal and process approaches	537

GUIDELINES FOR MANUSCRIPTS

Before finalizing your text for submission, please examine a recent issue of *Survey Methodology* (Vol. 32, No. 2 and onward) as a guide and note particularly the points below. Articles must be submitted in machine-readable form, preferably in Word. A pdf or paper copy may be required for formulas and figures.

1. Layout

- 1.1 Documents should be typed entirely double spaced with margins of at least 1½ inches on all sides.
- 1.2 The documents should be divided into numbered sections with suitable verbal titles.
- 1.3 The name (fully spelled out) and address of each author should be given as a footnote on the first page of the manuscript.
- 1.4 Acknowledgements should appear at the end of the text.
- 1.5 Any appendix should be placed after the acknowledgements but before the list of references.

2. Abstract

The manuscript should begin with an abstract consisting of one paragraph followed by three to six key words. Avoid mathematical expressions in the abstract.

3. Style

- 3.1 Avoid footnotes, abbreviations, and acronyms.
- 3.2 Mathematical symbols will be italicized unless specified otherwise except for functional symbols such as “exp(·)” and “log(·)”, *etc.*
- 3.3 Short formulae should be left in the text but everything in the text should fit in single spacing. Long and important equations should be separated from the text and numbered consecutively with arabic numerals on the right if they are to be referred to later.
- 3.4 Write fractions in the text using a solidus.
- 3.5 Distinguish between ambiguous characters, (*e.g.*, w, ω; o, O, 0; l, 1).
- 3.6 Italics are used for emphasis.

4. Figures and Tables

- 4.1 All figures and tables should be numbered consecutively with arabic numerals, with titles that are as self explanatory as possible, at the bottom for figures and at the top for tables.

5. References

- 5.1 References in the text should be cited with authors' names and the date of publication. If part of a reference is cited, indicate after the reference, *e.g.*, Cochran (1977, page 164).
- 5.2 The list of references at the end of the manuscript should be arranged alphabetically and for the same author chronologically. Distinguish publications of the same author in the same year by attaching a, b, c to the year of publication. Journal titles should not be abbreviated. Follow the same format used in recent issues.

6. Short Notes

- 6.1 Documents submitted for the short notes section must have a maximum of 3,000 words.

DIRECTIVES CONCERNANT LA PRÉSENTATION DES TEXTES

Avant de finaliser votre texte pour le soumettre, prière d'examiner un numéro récent de *Techniques d'enquête* (à partir du vol. 32, N° 2) et de noter les points ci-dessous. Les articles doivent être soumis sous forme de fichiers de traitement de texte, préférablement Word. Une version pdf ou papier pourrait être requise pour les formules et graphiques.

1.	Présentation	1.1 Les textes doivent être écrits à double interligne partout et avec des marges d'au moins 1 1/2 pouce tout autour. 1.2 Les textes doivent être divisés en sections numérotées portant des titres appropriés. 1.3 Le nom (écrit au long) et l'adresse de chaque auteur doivent figurer dans une note au bas de la première page du texte. 1.4 Les remerciements doivent paraître à la fin du texte. 1.5 Toute annexe doit suivre les remerciements mais précéder la bibliographie.
2.	Résumé	Le texte doit commencer par un résumé composé d'un paragraphe suivi de trois à six mots clés. Éviter les expressions mathématiques dans le résumé.
3.	Rédaction	3.1 Éviter les notes au bas des pages, les abréviations et les sigles. 3.2 Les symboles mathématiques seront imprimés en italique à moins d'une indication contraire, sauf pour les symboles fonctionnels comme $\exp(\cdot)$ et $\log(\cdot)$ etc. 3.3 Les formules courtes doivent figurer dans le texte principal, mais tous les caractères dans le texte doivent correspondre à un espace simple. Les équations longues et importantes doivent être séparées du texte principal et numérotées en ordre consécutif par un chiffre arabe à la droite si l'auteur y fait référence plus loin. 3.4 Écrire les fractions dans le texte à l'aide d'une barre oblique. 3.5 Distinguer clairement les caractères ambigus (comme w , ω ; o , O , 0 ; l , 1). 3.6 Les caractères italiques sont utilisés pour faire ressortir des mots.
4.	Figures et tableaux	4.1 Les figures et les tableaux doivent tous être numérotés en ordre consécutif avec des chiffres arabes et porter un titre aussi explicatif que possible (au bas des figures et en haut des tableaux).
5.	Bibliographie	5.1 Les références à d'autres travaux faites dans le texte doivent préciser le nom des auteurs et la date de publication. Si une partie d'un document est citée, indiquer laquelle après la référence. Exemple: Cochran (1977, page 164). 5.2 La bibliographie à la fin d'un texte doit être en ordre alphabétique et les titres d'un même auteur doivent être en ordre chronologique. Distinguer les publications d'un même auteur et d'une même année en ajoutant les lettres a, b, c, etc. à l'année de publication. Les titres de revues doivent être écrits au long. Suivre le modèle utilisé dans les numéros récents.
6.	Communications brèves	6.1 Les documents soumis pour la section des communications brèves doivent avoir au plus 3 000 mots.

CONTENTS

TABLE DES MATIÈRES

Volume 39, No. 3, September/septembre 2011

Special Issue: Special Issue in Honour of Jack Kalbfleisch and Jerry Lawless

Richard J. Cook and Grace Y. Yi	A special issue of CJS in honour of Jack Kalbfleisch and Jerry Lawless	385
Jiahua Chen and Pengfei Li	Tuning the EM-test for finite mixture models.....	389
Niels Keiding	Age-period-cohort analysis in the 1870s: Diagrams, stereograms, and the basic differential equation.....	405
Daeyoung Kim and Bruce G. Lindsay	Modal simulation and visualization in finite mixture models	421
Terry C.K. Lee, Leilei Zeng, Darby J.S. Thompson and C.B. Dean	Comparison of imputation methods for interval censored time-to-event data in joint modelling of tree growth and mortality.....	438
Ni Li, Do-Hwan Park, Jiansuo Sun and KyungMann Kim	Semiparametric transformation models for multivariate panel count data with dependent observation process	458
Karen McKeown and Nicholas P. Jewell	Current status observation of a three-state counting process with application to simultaneous accurate and diluted HIV test data.....	475
John M. Neuhaus and Charles E. McCulloch	The effect of misspecification of random effects distributions in clustered data settings with outcome-dependent sampling.....	488
Ross L. Prentice and Ying Huang	Measurement error modeling and nutritional epidemiology association analyses.....	498
Jing Qin and Biao Zhang	Optimal estimating functions in incomplete data and length biased sampling data problems	510
Alastair J. Scott and Chris J. Wild	Fitting regression models with response-biased samples.....	519
Yang Yang and Vijayan N. Nair	Parametric inference for time-to-failure in multi-state semi-Markov models: A comparison of marginal and process approaches	537

TABLE DES MATIÈRES

CONTENTS

Volume 39, No. 2, June/juin 2011

Alison L. Gibbs, Kevin J. Keen and Liqun Wang	Case studies in data analysis	181
Xiaoqing Niu, Pengfei Li and Peng Zhang	Testing homogeneity in a multivariate mixture model	218
Aleksey Min and Claudia Czado	Bayesian model selection for D-vine pair-copula constructions	239
Giampiero Marra and Rosalba Radice	Estimation of a semiparametric recursive bivariate probit model in the presence of endogeneity	259
Richard Charnigo and Cidambi Srinivasan	Self-consistent estimation of mean response functions and their derivatives	280
Seo Young Park and Yufeng Liu	Robust penalized logistic regression with truncated loss functions	300
Huazhen Lin, Ling Zhou, Heng Peng and Xiao-Hua Zhou	Selection and combination of biomarkers using ROC method for disease classification and prediction	324
Stefan H. Steiner, Nathaniel T. Stevens, Ryan Browne and R. Jock Mackay	Planning and analysis of measurement reliability studies	344
Shojaeddin Chenouri, Christopher G. Small and Thomas J. Farrar	Data depth-based nonparametric scale tests	356
Mingqiang Li, Liang Peng and Yongcheng Qi	Reduce computation in profile empirical likelihood method	370

JOURNAL OF OFFICIAL STATISTICS

An International Review Published by Statistics Sweden

JOS is a scholarly quarterly that specializes in statistical methodology and applications. Survey methodology and other issues pertinent to the production of statistics at national offices and other statistical organizations are emphasized. All manuscripts are rigorously reviewed by independent referees and members of the Editorial Board.

Contents

Volume 27, No. 3, 2011

A Unit-Error Theory for Register-Based Household Statistics	Li-Chun Zhang.....	415
Using Statistical Models for Sample Design of a Reinterview Program	Jianzhu Li, J. Michael Brick, Bac Tran, Phyllis Singer.....	433
Centre Sampling Technique in Foreign Migration Surveys: A Methodological Note	Gianluca Baio, Gian Carlo Blangiardo, Marta Blangiardo.....	451
Algorithms for Correcting Sign Errors and Rounding Errors in Business Survey Data	Sander Scholtus.....	467
Bayesian Estimation of Population Size via Linkage of Multivariate Normal Data Sets	Brunero Liseo, Andrea Tancredi.....	491
Random Group Variance Estimators for Survey Data with Random Hot Deck Imputation	Jun Shao, Qi Tang.....	507
Statistical Properties of Multiplicative Noise Masking for Confidentiality Protection	Tapan K. Nayak, Bimal Sinha, Laura Zayat.....	527
Book Review.....		545
In Other Journals.....		551

All inquiries about submissions and subscriptions should be directed to jos@scb.se

JOS is a scholarly quarterly that specializes in statistical methodology and applications. Survey methodology and other issues pertinent to the production of statistics at national offices and other statistical organizations are emphasized. All manuscripts are rigorously reviewed by independent referees and members of the Editorial Board.

JOURNAL OF OFFICIAL STATISTICS

An International Review Published by Statistics Sweden

Contents Volume 27, No. 2, 2011

Preface	151
Annelies Blom, Frauke Kreuter	
Proxy Pattern-Mixture Analysis for Survey Nonresponse	153
Rebecca R. Andridge, Roderick J.A. Little	
Imputation and Estimation under Nonignorable Nonresponse in Household Surveys with Missing Covariate Information	181
Danny Pfeffermann, Anna Sikov	
An Analysis of Nonignorable Nonresponse to Income in a Survey with a Rotating Panel Design	211
Caterina Giusti, Roderick J.A. Little	
Indicators for Monitoring and Improving Representativeness of Response	231
Barry Schouten, Natalie Shlomo, Chris Skinner	
A Pseudo-GEE Approach to Analyzing Longitudinal Surveys under Imputation for Missing Responses	255
Iván A. Carrillo, Jiahua Chen, Changbao Wu	
Effects of Increasing the Incentive Size in a Longitudinal Study	279
Willard L. Rodgers	
Attrition in the Swiss Household Panel: Is Change Associated with Drop-out?	301
Martieke Voorpostel, Oliver Lipps	
Keeping Track of Panel Members: An Experimental Test of a Between-Wave Contact Strategy	319
Katherine A. McGonagle, Mick P. Couper, Robert F. Schoeni	
Using Paradata and Other Auxiliary Data to Examine Mode Switch Nonresponse in a "Recruit-and-Switch" Telephone Survey	339
Joseph W. Sakshaug, Frauke Kreuter	
Interviewer Effects on Nonresponse in the European Social Survey	359
Annelies G. Blom, Edith D. de Leeuw, Joop J. Hox	
Toward a Benefit-Cost Theory of Survey Participation: Evidence, Further Tests, and Implications	379
Eleanor Singer	
Applying Motivation Theory to Achieve Increased Response Rates, Respondent Satisfaction and Data Quality	393
Marika Wenneberg, Andreas Persson, Helle Noorlind Brage, Tommy Svensson, Margareta Kristenson	

All inquiries about submissions and subscriptions should be directed to jos@scb.se

Membres du comité de sélection de l'article Waksberg (2011-2012)

Mary Thompson, *University of Waterloo* (Présidente)

J.N.K. Rao, *Carleton University*

Steve Heeringa, *University of Michigan*

Cynthia Clark, *USDA*

Présidents précédents :

Graham Kalton (1999 - 2001)

Chris Skinner (2001 - 2002)

David A. Binder (2002 - 2003)

J. Michael Brick (2003 - 2004)

David R. Bellhouse (2004 - 2005)

Gordon Brackstone (2005 - 2006)

Sharon Lohr (2006 - 2007)

Robert Groves (2007 - 2008)

Leyla Mojadjer (2008 - 2009)

Daniel Kasprzyk (2009 - 2010)

Elizabeth A. Martin (2010 - 2011)

ANNONCES

Demande de candidatures pour le prix Waksberg 2013

La revue *Techniques d'enquête* a mis sur pied une série de communications sollicitées en l'honneur de Joseph Waksberg, qui a fait de nombreuses contributions importantes à la méthodologie d'enquête. Chaque année, un éminent chercheur est choisi pour rédiger un article pour la série de communications sollicitées de Waksberg. L'article examine les progrès et l'état actuel d'un thème important dans le domaine de la méthodologie d'enquête et reflète l'agencement de théorie et de pratique caractéristique des travaux de Waksberg.

L'auteur reçoit une prime en argent qui provient d'une bourse de Westat. L'article sera publié dans un numéro futur de *Techniques d'enquête*.

L'auteur de l'article Waksberg de 2013 sera sélectionné par un comité de quatre personnes désignées par *Techniques d'enquête* et l'American Statistical Association. Les candidatures ou les suggestions de sujets doivent être envoyées avant le 28 février 2012 à la présidente du comité Mary Thompson (methompson@uwaterloo.ca).

Les gagnants et articles précédents du prix Waksberg sont

- 2001 Gad Nathan, « Méthodes de téléenquêtes applicables aux enquêtes-ménages – Revue et réflexions sur l'avenir », *Techniques d'enquête*, vol. 27, 1, 7-34.
- 2002 Wayne A. Fuller, « Estimation par régression appliquée à l'échantillonnage », *Techniques d'enquête*, vol. 28, 1, 5-25.
- 2003 David Holt, « Enjeux méthodologiques de l'élaboration et de l'utilisation d'indicateurs statistiques pour des fins de comparaisons internationales », *Techniques d'enquête*, vol. 29, 1, 5-19.
- 2004 Norman M. Bradburn, « Comprendre le processus de question et réponse », *Techniques d'enquête*, vol. 30, 1, 5-16.
- 2005 J.N.K. Rao, « Évaluation de l'interaction entre la théorie et la pratique des enquêtes par sondage », *Techniques d'enquête*, vol. 31, 2, 127-151.
- 2006 Alastair Scott, « Études cas-témoins basées sur la population », *Techniques d'enquête*, vol. 32, 2, 137-147.
- 2007 Carl-Erik Särndal, « La méthode de calage dans la théorie et la pratique des enquêtes », *Techniques d'enquête*, vol. 33, 2, 113-135.
- 2008 Mary E. Thompson, « Enquêtes internationales : motifs et méthodologies », *Techniques d'enquête*, vol. 34, 2, 145-157.
- 2009 Graham Kalton, « Méthodes de suréchantillonnage des sous-populations rares dans les enquêtes sociales », *Techniques d'enquête*, vol. 35, 2, 133-152.
- 2010 Ivan P. Fellegi, « L'organisation de la méthodologie statistique et de la recherche méthodologique dans les bureaux nationaux de la statistique », *Techniques d'enquête*, vol. 36, 2, 131-139.
- 2011 Danny Pfeffermann, « Modélisation des données d'enquêtes complexes : Pourquoi les modéliser ? Pourquoi est-ce un problème ? Comment le résoudre ? », *Techniques d'enquête*, vol. 37, 2, 123-146.
- 2012 Lars Lyberg, Sujet de l'article à l'étude.

ou plus durant l'année 2011.

J.M. Aitkin, <i>University of Melbourne</i>	J.-F. Beaumont, <i>Statistique Canada</i>	W. Bell, <i>U.S. Census Bureau</i>	E. Berg, <i>Iowa State University</i>	Y. Berger, <i>University of Southampton</i>	P. Biemer, <i>Research Triangle Institute</i>	C. Bocci, <i>Statistique Canada</i>	J. van den Brakel, <i>Statistics Netherlands</i>	J.M. Brick, <i>Westat Inc.</i>	J. Bushery, <i>U.S. Census Bureau</i>	R. Cantwell, <i>U.S. Bureau of the Census</i>	S. Eckman, <i>Institute for Employment Research</i>	A. Eideh, <i>Al-Quds University</i>	M. Elliott, <i>University of Michigan</i>	J.L. Eitinge, <i>U.S. Bureau of Labor Statistics</i>	V. Esteveao, <i>Statistique Canada</i>	G. Forsman, <i>Swedish Transport Administration</i>	W.A. Fuller, <i>Iowa State University</i>	D. Haziza, <i>Université de Montréal</i>	Y. He, <i>Harvard Medical School</i>	S. Heeringa, <i>University of Michigan</i>	M. Hidiroglou, <i>Statistique Canada</i>	B. Hultiger, <i>University of Applied Sciences Northwestern Switzerland</i>	P.M. Joyce, <i>U.S. Bureau of Census</i>	D. Judkins, <i>Westat Inc.</i>	D. Kasprzyk, <i>NORC at the University of Chicago</i>	J.-K. Kim, <i>Iowa State University</i>	P. Kott, <i>Research Triangle Institute</i>	T. Krenzke, <i>Westat Inc.</i>	P. Lahiri, <i>JPSM, University of Maryland</i>	P. Lavallée, <i>Statistique Canada</i>	H. Lee, <i>Westat Inc.</i>	J. Legg, <i>Amgen Inc., États-Unis</i>	J. Li, <i>Westat Inc.</i>	S. Lohr, <i>Arizona State University</i>	E. López Escobar, <i>University of Southampton</i>	P. Lynn, <i>University of Essex</i>	D.J. Mailec, <i>National Center for Health Statistics</i>	H. Manel, <i>Statistique Canada</i>	K. Miller, <i>U.S. National Center for Health Statistics</i>	J.M. Montaquila, <i>Westat Inc.</i>	F. Moura, <i>Universidade Federal do Rio de Janeiro</i>	T. Mulcahy, <i>National Opinion Research Center</i>	J.F. Muñoz Rosas, <i>University of Granada</i>	B. Nandram, <i>WPI</i>	G. Nathan, <i>Hebrew University</i>	D. Nelson, <i>Center for Chronic Disease Outcomes Research</i>	S. Oman, <i>Hebrew University</i>	J. Opsomer, <i>Colorado State University</i>	J.L. Parsons, <i>USDA, National Agricultural Statistics Service</i>	Z. Patak, <i>Statistique Canada</i>	D. Pfeffermann, <i>Hebrew University</i>	C. Poirier, <i>Statistique Canada</i>	N.G.N. Prasad, <i>University of Alberta</i>	J. Preston, <i>AB5</i>	S. Raabe-Hesketh, <i>UC Berkeley</i>	J.N.K. Rao, <i>Carleton University</i>	J. Reiter, <i>Duke University</i>	L.-P. Rivest, <i>Université Laval</i>	S. Rubin-Bleuer, <i>Statistique Canada</i>	N. Schenker, <i>National Center for Health Statistics</i>	F.J. Scheuren, <i>National Opinion Research Center</i>	B. Schouten, <i>Statistics Netherlands</i>	R. Sigmán, <i>Westat Inc.</i>	P. do N. Silva, <i>Escola Nacional de Ciências Estatísticas</i>	M. Sinclair, <i>NORC at the University of Chicago</i>	A. Singh, <i>National Opinion Research Center</i>	B. Skalland, <i>National Opinion Research Center</i>	E. Sluid, <i>University of Maryland</i>	P. Smith, <i>Office for National Statistics</i>	E. Stasny, <i>Ohio State University</i>	D. Steel, <i>University of Wollongong</i>	L. Stokes, <i>Southern Methodist University</i>	M. Sverchikov, <i>Bureau of Labor Statistics</i>	M. Thompson, <i>University of Waterloo</i>	C. Tucker, <i>Consultant</i>	N. Tzavidis, <i>University of Southampton</i>	B. van der Klaauw, <i>University of Amsterdam</i>	J. van der Laan, <i>Statistics Netherlands</i>	V.J. Verma, <i>Università degli Studi di Siena</i>	J. Wang, <i>Hewlett-Packard Labs</i>	Z. Wang, <i>Sir Wilfrid Laurier University</i>	K.M. Wolter, <i>National Opinion Research Center</i>	J. Wood, <i>Office for National Statistics</i>	C. Wu, <i>University of Waterloo</i>	Y. You, <i>Statistique Canada</i>	C. Yu, <i>Iowa State University</i>	W. Yung, <i>Statistique Canada</i>	A. Zaslavsky, <i>Harvard Medical School</i>	L.-C. Zhang, <i>Statistics Norway</i>
---	---	------------------------------------	---------------------------------------	---	---	-------------------------------------	--	--------------------------------	---------------------------------------	---	---	-------------------------------------	---	--	--	---	---	--	--------------------------------------	--	--	---	--	--------------------------------	---	---	---	--------------------------------	--	--	----------------------------	--	---------------------------	--	--	-------------------------------------	---	-------------------------------------	--	-------------------------------------	---	---	--	------------------------	-------------------------------------	--	-----------------------------------	--	---	-------------------------------------	--	---------------------------------------	---	------------------------	--------------------------------------	--	-----------------------------------	---------------------------------------	--	---	--	--	-------------------------------	---	---	---	--	---	---	---	---	---	--	--	------------------------------	---	---	--	--	--------------------------------------	--	--	--	--------------------------------------	-----------------------------------	-------------------------------------	------------------------------------	---	---------------------------------------

Nous remercions également ceux qui ont contribué à la production des numéros de la revue pour 2011 : Céline Ethier de la Division de la recherche et de l'innovation en statistique, Christine Cousineau et Teresa Jewell de la Division des méthodes d'enquêtes auprès des ménages, Nick Budko et Sophie Chartier de la Division des méthodes d'enquêtes auprès des ménages, Dominique Lavoie de la Division des méthodes d'enquêtes sociales, Mathew Belyea, Louise Demers, Anne-Marie Fleury, Roberto Guido, Lydia Kokline, Liliane Lanoie, Darqaise Pellerin, Joseph Prince et Fadi Salibi de la Division de la diffusion.

approximativement équivalente à l'échantillonnage équilibré.

5. Conclusion

Les méthodes décrites dans les trois exposés sont remarquablement complémentaires. Les plans d'échantillonnage adaptés permettent d'obtenir des échantillons aléatoires, statistiquement valides, pour des populations habituellement difficiles à échantillonner efficacement. Très peu d'information est nécessaire pour tirer ce genre d'échantillon, mais de nombreux efforts doivent être faits durant la collecte des données afin de découvrir et de suivre les « liens » entre les éléments de la population et de tirer des échantillons successifs. En revanche, l'échantillonnage équilibré est utile quand des renseignements très détaillés sont disponibles dans la base de sondage et, dans cette situation, il permet d'obtenir des plans d'échantillonnage hautement personnalisés et efficaces. Une fois qu'un échantillon équilibré est tiré, la collecte des données peut se poursuivre de la même façon que dans les enquêtes habituelles. L'échantillonnage à bases de sondage multiples couvre un cas intermédiaire, en ce sens qu'aucune bonne base de sondage unique n'existe, mais que plusieurs bases de sondage partielles sont utilisées pour « compenser » leur faiblesses réciproques. Des échantillons distincts sont tirés de chaque base de sondage et la collecte des données se fait comme d'habitude, sauf qu'il est nécessaire de déterminer à

quelle(s) base(s) de sondage chaque unité échantillonnée appartient.

Conjuguées aux approches existantes déjà mises en œuvre, ces trois nouvelles méthodes d'échantillonnage pourraient accroître fortement la souplesse avec laquelle des échantillons peuvent être adaptés à des applications particulières, afin de réduire les coûts d'enquête et d'augmenter la précision des estimateurs.

Bibliographie

Deville, J.-C., et Tillé, Y. (2005). Variance approximation under balanced sampling. *Journal of Statistical Planning and Inference*, 2, 569-591.

Fuller, W.A. (2009). Some design properties of a rejective sampling procedure. *Biometrika*, 96(4), 933-944.

Kott, P.S. (2001). The delete-a-group jackknife. *Journal of Official Statistics*, 17, 521-526.

Lohr, S. (2007). Recent developments in multiple frame surveys. *Dans ASA Proceedings of the Joint Statistical Meetings*, American Statistical Association, 3257-3264.

Skinner, C.J., et Rao, J.N.K. (1996). Estimation in dual frame surveys with complex designs. *Journal of the American Statistical Association*, 91, 349-356.

Thompson, S.K. (2006). Adaptive web sampling. *Biometrics*, 62, 1224-1234.

probabilités d'inclusion postulées et satisfaisant exactement les contraintes d'équilibrage ne peut être trouvé, l'algorithme essayera de trouver une solution qui satisfait d'au moins une partie des contraintes. La méthode du cube requiert que les variables d'équilibrage x_i soient connues pour tous les éléments de la population. Selon le contexte de l'enquête, cette exigence pourrait représenter une limite importante de l'applicabilité de l'échantillonnage équilibré. Même si l'équilibrage sur les variables auxiliaires au niveau de la population est effectué à l'étape de l'élaboration du plan, il paraît probable qu'en pratique, le calage et d'autres corrections de la pondération, telles que celles de la non-réponse, soit souvent requis. En fait, selon Tillé, la combinaison de l'équilibrage et du calage constitue la stratégie la plus efficace (voir la section 7.4 de l'article de Tillé dans le présent numéro). Toutefois, les propriétés théoriques des estimateurs qui sont à la fois équilibrés et calés n'ont pas encore été établies complètement.

Bien que l'échantillonnage équilibré permette de maintenir les probabilités d'inclusion des éléments dans la population, il est clair que l'existence de contraintes d'équilibrage affecte les probabilités d'inclusion *conjointes* et donc la variance des estimateurs. Ce sujet est abordé à la section 6 de l'article de Tillé. Deville et Tillé (2005) ont montré que, dans certaines conditions, la variance des estimateurs sous échantillonnage équilibré peut être approximée par une variance de type linéarisation, qui dépend des résidus d'une régression linéaire des variables étudiées sur les variables d'équilibrage. Quoiqu'il s'agisse d'un résultat important et utile, il ne mène pas à une approche d'estimation de la variance convenant à toutes les applications d'enquête. L'un des problèmes est que l'estimation de la variance fondée sur cette méthode requiert l'accès aux variables d'équilibrage pour tous les participants à l'enquête et que ces variables pourraient ne pas être diffusées publiquement dans l'enquête de données d'enquête. Dans ce contexte, une méthode de rééchantillonnage pourrait être particulièrement séduisante, parce qu'elle ne nécessiterait pas la diffusion de ces variables. Cependant, aucune méthode de ce genre n'est disponible à l'heure actuelle.

L'échantillonnage équilibré présente des liens étroits avec l'échantillonnage *réjectif*, dont les objectifs sont les mêmes. Dans l'échantillonnage *réjectif*, un échantillon est tiré avec des probabilités d'inclusion préspecifiées, et l'échantillon est accepté ou rejeté selon qu'il est compris ou non dans une fourchette de tolérance donnée d'une contrainte d'équilibrage. Si l'échantillon est rejeté, la procédure est répétée jusqu'à ce que soit trouvé un échantillon qui se trouve dans la fourchette de tolérance. L'échantillonnage *réjectif* existe de longue date, mais Fuller (2009) a décrit une certaine théorie asymptotique montrant qu'asymptotiquement, sa version de l'échantillonnage *réjectif* était

cube, qu'il a élaborée en tant qu'algorithme pratique pour le tirage d'échantillons équilibrés. Les objectifs des plans d'échantillonnage équilibré consistent à maintenir la représentation de la structure de la population dans l'échantillon (d'où le terme « équilibré ») et à améliorer l'efficacité des estimateurs d'après des données d'enquête. Aujourd'hui, la stratification est le principal outil qu'utilisent la plupart des statisticiens d'enquête en vue de réaliser ces deux objectifs. La stratification permet d'atteindre l'équilibre en forçant la composition de l'échantillon à concorder avec la répartition entre les strates et accroît l'efficacité des estimateurs en éliminant la composante de la variance due aux différences entre strates. L'échantillonnage systématique est également utilisé pour atteindre ces objectifs, le plus souvent dans le contexte des enquêtes sur les ressources naturelles. Dans ce cas, la composition de l'échantillon correspond exactement à la composition de la population pour les variables de tirage correspond approximativement pour toute variable corrélée à la variable de tirage. Un gain d'efficacité est réalisé parce que les moments des variables d'intérêt dans l'échantillon concordent (approximativement) avec les moments dans la population. Bien que les deux approches soient utilisées à grande échelle et donnent de bons résultats, elles manquent de souplesse. La stratification requiert souvent de diviser la population en « cellules » définies par l'intersection des variables de stratification, ce qui peut donner lieu à une prolifération de petites cellules correspondant à de petites tailles d'échantillon. L'échantillonnage systématique est une forme hautement contraignante d'échantillonnage qui offre une souplesse limitée en ce qui concerne la construction de l'échantillon et qui pose aussi le problème de l'absence d'un estimateur de variance fondé sur le plan de sondage.

L'échantillonnage équilibré peut être considéré comme une généralisation de la stratification. Sous cette interprétation, les échantillons stratifiés sont tirés avec des probabilités d'inclusion données pour tous les éléments de la population, mais sous la contrainte de la taille d'échantillon dans chaque strate. Dans l'échantillonnage équilibré, les contraintes de stratification sont remplacées par des contraintes de la forme $\sum_i x_i / \pi_i = \sum_j x_j$, où x_i est un vecteur de variables d'équilibrage. Quand les x_i sont des indicateurs de strate, l'échantillonnage équilibré coïncide avec la stratification, mais toute variable catégorique ou continue (ou une combinaison de celles-ci) peut être utilisée, ce qui donne une grande souplesse pour la construction de l'échantillon.

Comme il est mentionné plus haut, la méthode du cube est un algorithme qui permet de tirer des échantillons équilibrés étant donné un ensemble de probabilités d'inclusion et de contraintes. Si des échantillons exactement équilibrés existent dans la population, l'algorithme essayera de sélectionner l'un d'eux. Si aucun échantillon ayant les

méthodes de rééchantillonnage. Dans le cas de l'approche par linéarisation, une question pratique importante tient au fait qu'il faut pouvoir déterminer à quelle base de sondage appartiennent chaque élément de l'échantillon, car la variance doit être estimée séparément dans chaque base de sondage. L'organisme d'enquête qui produit les données pourrait juger cette situation indésirable, pour des raisons de confidentialité des données. Dans le cas des méthodes de rééchantillonnage, telles que le jackknife et le bootstrap, l'organisme d'enquête peut créer des ensembles de poids de rééchantillonnage qui ne requièrent pas que l'on divulgue aux utilisateurs des données à quelle base de sondage appartiennent les divers éléments de l'échantillon. Lohr (2007) recommande l'approche du *bootstrap combiné* pour l'inférence sous échantillonnage à bases de sondage multiples. La méthode du *jackknife groupé* de Kott (2001) pourrait également être considérée comme une autre solution.

La mise en œuvre d'enquêtes à bases de sondage multiples est parfois plus difficile que celle d'enquêtes à base de sondage unique. Comme il est mentionné à la section 5 de l'article de Lohr, il faut être conscient du risque accru d'erreurs non dues à l'échantillonnage, surtout si les modes ou protocoles de collecte des données varient d'une base de sondage à l'autre. Par exemple, les éléments échantillonnés dans une base de sondage reçoivent une lettre de présentation envoyée à l'avance, tandis que ceux d'une autre base de sondage reçoivent un « appel direct » à cause du manque de renseignements sur l'adresse. Il se peut aussi que les caractéristiques de la non-réponse diffèrent selon la base de sondage, de sorte que des corrections distinctes sont nécessaires. Enfin, dans de nombreux cas, les éléments présents dans les diverses bases de sondage pourraient avoir des caractéristiques différentes (par exemple, fermes biologiques membres d'une association nationale d'entrepreneurs de type biologique vs celles qui n'en sont pas membres). Dans tous ces cas, il convient de faire attention aux effets propres à la base de sondage et de construire prudemment les pondérations afin d'obtenir des estimateurs pour données d'enquête valides. Par ailleurs, l'existence de multiples bases de sondage offre la possibilité de mesurer les erreurs non dues à l'échantillonnage, parce qu'elles fournissent des échantillons multiples d'une même population. Par exemple, il pourrait être utile d'effectuer des « appels directs » auprès d'une partie des éléments sélectionnés dans la base de sondage contenant des adresses pour évaluer les effets de mode.

4. Échantillonnage équilibré par la méthode du cube

Dans son exposé, le professeur Tillé a énoncé les fondements de l'échantillonnage équilibré et décrit la méthode du

d'enquêtes auprès de ce genre de sous-populations. L'approche à bases de sondage multiples pourrait également être utile pour une enquête auprès de la population générale comme moyen d'accroître la taille de l'échantillon dans certaines sous-populations présentant un intérêt particulier. Par exemple, dans une enquête générale sur les fermes, on pourrait souhaiter produire des estimations pour les fermes de type biologique, qui ne représentent qu'une faible fraction des fermes, mais dont bon nombres figurent dans les répertoires d'entreprises de type biologique. La section 1 de l'article de Lohr publié dans le présent numéro donne plusieurs autres exemples du vaste champ d'application des enquêtes à bases de sondage multiples.

Comme je l'ai mentionné plus haut, les méthodes d'estimation comprennent la construction d'estimateurs pour la sous-population contenue dans l'intersection des bases de sondage, ce qui requiert le choix d'une méthode de pondération pour les estimateurs obtenus d'après les différentes bases de sondage. Les méthodes de pondération qui s'appuient sur l'estimation de la précision de ces estimateurs pourraient être privilégiées dans une perspective d'efficacité. Cependant, leur mise en œuvre en pratique pose quelques problèmes, parce que les poids résultants peuvent varier pour diverses variables de l'enquête. Des approches plus commodées consistent à renoncer à une certaine efficacité afin de pouvoir utiliser les mêmes poids pour toutes les variables de l'enquête, une caractéristique soutenue à plusieurs reprises dans l'article de Lohr publié dans le présent numéro. La méthode du *pseudo-maximum de vraisemblance* (PMV) de Skinner et Rao (1996), qui produit un ensemble unique de poids, est recommandée par Lohr comme méthode privilégiée pour les enquêtes uniques, tandis qu'une approche à poids fixe plus simple est préférable pour les enquêtes longitudinales.

Bien que la méthodologie de base pour la construction d'estimateurs fondés sur le plan pour l'échantillonnage à bases de sondage multiples soit établie aujourd'hui, il est nécessaire de poursuivre l'étude d'approches pour appliquer le calage et la correction de la non-réponse dans ce contexte. Comme il est possible d'appliquer ces corrections au niveau de la base de sondage individuelle, au niveau de la population ou aux deux niveaux (selon l'information auxiliaire disponible), une étude des propriétés des estimateurs sous ces divers scénarios serait fort utile et devrait servir à élaborer des lignes directrices à l'intention des praticiens des sondages. La section 3 de l'article de Lohr dans le présent numéro offre une discussion de quelques-uns des premiers résultats dans ce domaine.

La section 4.2 de l'article de Lohr est consacrée à l'examen des méthodes d'estimation de la variance des estimateurs pour bases de sondage multiples qui ont été élaborées, y compris les méthodes de linéarisation et les

fondée sur la randomisation et l'échantillonnage dans les bases de sondage individuelles se fait selon la même méthodologie que l'échantillonnage « classique » dans une seule base de sondage. Des méthodes d'estimation entièrement fondées sur le plan de sondage existent pour l'échantillonnage à bases de sondage multiples et plusieurs d'entre elles peuvent être déployées facilement dans le contexte des enquêtes à grande échelle dans lesquelles un ensemble de données pondérées est le principal produit (voir plus bas). La caractéristique principale de toutes les méthodes d'estimation est l'estimation du chevauchement des bases de sondage, qui est habituellement inconnu, mais qui doit être pris en compte. Pour l'estimer, on construit, pour chaque base de sondage, des estimateurs fondés sur le plan pour la ou les sous-populations d'éléments qui se retrouvent aussi dans la ou les autres bases de sondage. Les estimateurs des caractéristiques de la ou des intersections entre les bases de sondage doivent alors être combinés sur l'ensemble des bases de sondage. Les méthodes existantes diffèrent quant à la façon de combiner ces estimateurs, les plus simples utilisant des moyennes pondérées par la taille d'échantillon et les plus complexes, des estimateurs pondérés par des estimations de la précision des estimateurs individuels.

sont l'échantillonnage par composition aléatoire, où une fraction croissante de la population ne peut pas être rejointe au moyen d'un numéro de téléphone fixe, les enquêtes auprès de professionnels ou d'entreprises pour lesquelles des listes partielles peuvent être obtenues auprès de fournisseurs ou d'organismes professionnels. Les enquêtes sur des populations rares qui existent au sein de la population plus générale sont d'autres situations dans lesquelles l'échantillonnage à bases de sondage multiples pourrait être appliqué. Une base de sondage globale existe pour la population, mais la sélection des répondants pour savoir s'ils appartiennent à la sous-population d'intérêt est longue et coûteuse. Une autre base de sondage contenant une portion nettement plus élevée d'éléments de la sous-population d'intérêt est parfois disponible, mais si la couverture de cette base de sondage est incomplète, l'organisme chargé de l'enquête pourrait ne pas vouloir s'en servir de crainte de ne pas obtenir un échantillon valide. La combinaison de cette base de sondage de rechange, mais incomplète, de la sous-population avec la base de sondage complète, de l'ensemble de la population pourrait être à la fois rentable et défendable du point de vue statistique. Les enquêtes sur la chasse et la pêche, pour lesquelles il existe souvent une liste des permis octroyés qui peut être incomplète ou non à jour, sont des exemples

multiples chevauchantes

Statistique Canada, N° 12-001-X au catalogue

suréchantillonner les UPE que l'on s'attend à contenir une fraction importante des groupes d'intérêt. Une enquête sur les hommes afro-américains courant le risque d'avoir la maladie de Parkinson dans laquelle pourraient être suréchantillonnées les secteurs de recensement comptant une échantillon élevée de la population afro-américaine en est un exemple. Un autre plan d'échantillonnage qui peut être utile dans ce contexte est l'échantillonnage à plusieurs phases. Dans ce cas, la première phase d'échantillonnage est utilisée comme échantillon de sélection ou comme un moyen de recueillir de l'information auxiliaire pertinente, tandis que les phases subséquentes visent à obtenir les données d'enquête d'intérêt. L'Agicultural Resource Management Survey (ARMS) (menée par le USDA) suit ce genre de plan. Un échantillon de toutes les fermes est sélectionné à la phase 1, en vue de recueillir des données sur les caractéristiques des fermes pour l'année de référence de l'enquête. Durant les phases ultérieures, on procède à la sélection d'échantillons ciblés en se basant sur les produits d'intérêt (par exemple, produits laitiers, blé, etc.). Une troisième approche d'échantillonnage parfois utile pour obtenir des échantillons de (sous-)populations rares est l'échantillonnage à bases multiples. Le principe qui sous-tend l'échantillonnage à bases multiples est la combinaison de plusieurs bases de sondage ayant différentes caractéristiques de couverture, par exemple une « bonne » base de sondage contenant une grande proportion des éléments d'intérêt, mais pouvant éventuellement être incomplète, et une « mauvaise » base de sondage qui est complète, mais ne contient qu'une faible proportion des éléments d'intérêt. Par exemple, une enquête auprès des sociétés d'une industrie particulière pourrait être réalisée en se servant d'une liste des membres de groupes d'industries comme « bonne » base de sondage et d'une liste générale des sociétés comme « mauvaise » base de sondage. Pour un examen plus approfondi de l'échantillonnage à bases de sondage multiples, voir la section 3 qui suit.

Comparativement à ces trois plans d'échantillonnage, l'échantillonnage adaptatif est plus souple et permet un contrôle plus fin du nombre et des caractéristiques des éléments qui sont inclus dans l'échantillon, ce qui, souvent, accroît l'efficacité et/ou réduit le coût. Un inconvénient de l'échantillonnage adaptatif est qu'il faut recueillir l'information sur les liens entre les éléments, ce qui peut augmenter le fardeau de réponse et le coût de la collecte, et éventuellement poser des problèmes de confidentialité.

Comme l'échantillonnage adaptatif s'appuie fréquemment sur des « liens » entre éléments afin de définir les probabilités de sélection conditionnelles dans les vagues d'échantillonnage, il convient aussi particulièrement bien pour les enquêtes qui visent à étudier les liens entre les

un estimateur sans biais de μ_N est donné par $z_k = \sum_{s_k=1}^{K-k-1} y_i + \sum_{s_k}^K y_i / q_{ki}$, où les q_{ki} sont les probabilités d'inclusion conditionnelles pour la vague k (voir Thompson (2006) pour des précisions sur la construction des q_{ki} , et la section 2.4 de l'article de Thompson dans le présent numéro pour des exemples spécifiques). En posant que $\hat{\mu}_r = K^{-1} \sum_{k=1}^K z_k$, un estimateur sans biais de μ_N s'obtient sous la forme $\hat{\mu} = w\hat{\mu}_0 + (1-w)\hat{\mu}_r$, qui est une combinaison linéaire de l'estimateur initial et de la moyenne des estimateurs subséquents. L'estimateur $\hat{\mu}$ est sans biais sous le plan, mais il dépend de l'ordre des vagues de sélection de l'échantillon. Un estimateur plus précis peut être obtenu en calculant la moyenne sur les divers ordres dans lesquels un échantillon aurait pu être obtenu. Pour les petites tailles d'échantillon, une expression explicite existe pour cet estimateur plus efficace, mais en général, il doit être approximé par échantillonnage répété à partir d'une chaîne de Markov définie de manière appropriée, puis par calcul de la moyenne des échantillons. Les méthodes exactes d'établissement de la chaîne et de tirage des échantillons sont décrites dans Thompson (2006), qui discute également de l'estimation de la variance de l'estimateur résultant.

L'un des principaux avantages des plans d'échantillonnage adaptatif tient au fait qu'ils permettent à l'organisme chargé des enquêtes de concentrer l'échantillon sur les parties de la population présentant un intérêt. Cela est particulièrement utile dans les situations où certains éléments d'intérêt sont relativement rares et qu'ils ne peuvent pas être identifiés a priori dans une base de sondage. Les enquêtes sur la chasse et la pêche, les nouveaux immigrants, les enfants scolarisés à domicile et les propriétaires d'entreprises familiales en sont des exemples. Dans chacun de ces cas, les éléments sont assez « dispersés » dans la population et aucune base de sondage complète n'est généralement disponible. Cependant, il est probable que les individus qui font partie de cette population seront capables de fournir des renseignements sur d'autres individus, de sorte que des liens peuvent être identifiés et échantillonnés au cours de différentes vagues d'échantillonnage adaptatif. Notons que l'échantillonnage adaptatif peut également être utilisé quand ces types d'éléments rares font partie d'une sous-population d'intérêt dans une enquête auprès d'une population plus grande et non rare. Par exemple, une enquête sur les écoliers pourrait inclure une strate d'enfants scolarisés à domicile. Rejoindre des (sous-)populations relativement rares est un défi fréquent dans les enquêtes, et un certain nombre de méthodes sont régulièrement déployées pour résoudre ce problème. Dans le contexte des enquêtes-ménages, le plan d'échantillonnage sans doute le plus fréquent est l'échantillonnage stratifié à plusieurs degrés. Dans la mesure où de l'information auxiliaire pertinente au niveau de l'UPE est disponible, l'organisme chargé de l'enquête peut

Plans d'échantillonnage novateurs : discussion de trois communications présentées au U.S. Census Bureau

Jean Opsomer¹

1. Introduction

Outre son rôle dans la collecte des données du recensement décennal des États-Unis, le U.S. Census Bureau est l'un des plus grands organismes de collecte de données d'enquête au monde. Les deux outils statistiques qu'il utilise principalement pour concevoir ses enquêtes sont la stratification et l'échantillonnage à plusieurs degrés. Mis en œuvre avec succès durant les années 1940, ces outils ont continué d'être adaptés et perfectionnés depuis.

Bien que cette approche générale d'échantillonnage ait été très fructueuse, la hausse des coûts d'enquête, la diminution des taux de réponse et l'existence de nouveaux problèmes de couverture des bases de sondage (surtout dans le cas des enquêtes téléphoniques) suscitent de plus en plus d'inquiétudes. Parallèlement, les progrès en ce qui concerne les méthodes de collecte des données, les nouvelles sources de données et les outils informatiques permettent d'envisager des plans d'enquête qui n'auraient pas été possibles auparavant. Dans le cadre du programme de remaniement entrepris en 2010, le Census Bureau a demandé à des chercheurs universitaires éminents de donner leur avis sur des méthodes d'échantillonnage novatrices, en vue de commencer à explorer de nouvelles approches possibles de conception de ses enquêtes. Ainsi, les professeurs Steve Thompson (Simon Fraser University), Sharon Lohr (Arizona State University) et Yves Tillé (Université de Neufchâtel) ont été invités à donner des exposés d'ensemble sur certains plans d'échantillonnage qu'ils ont élaborés. J'ai été invité à offrir une discussion sur chacun de ces exposés.

Dans les trois sections qui suivent, je résumerai mes commentaires sur chacun des exposés. Mes objectifs, en formulant ces commentaires, étaient de mettre en relief les aspects les plus importants des méthodes d'échantillonnage présentées, de discuter de quelques possibilités importantes de les utiliser dans le contexte de l'échantillonnage des ménages et de cerner les difficultés éventuelles de mise en œuvre.

2. Sondage par réseaux, échantillonnage adaptatif et échantillonnage dans l'espace

L'exposé du professeur Thompson portait sur une catégorie générale de plans d'échantillonnage qui englobent

l'échantillonnage adaptatif en grappes, le sondage par réseaux et l'échantillonnage en ligne adaptatif. Dans la suite du présent exposé, à moins de faire référence à un plan particulier dans cette classe, je donnerai à ces plans le nom d'« échantillonnage adaptatif ». Un avantage important de l'échantillonnage adaptatif tient au fait qu'il intègre certaines caractéristiques des approches d'échantillonnage « de commodité », telles que l'échantillonnage boule de neige, y compris le fait de s'appuyer moins sur une base de sondage et la capacité de cibler l'échantillonnage sur des parties de la population présentant un intérêt particulier. Cependant, contrairement à l'échantillonnage de commodité, l'échantillonnage adaptatif demeure fermement fondé sur un plan d'échantillonnage, au sens qu'il permet l'estimation et l'inférence de la population finie selon la randomisation.

Dans les procédures d'échantillonnage adaptatif, un échantillon initial est tiré conformément à un plan d'échantillonnage probabiliste $p_0(s_0)$. En fonction des caractéristiques des éléments compris dans s_0 (par exemple, présence/absence des caractéristiques d'intérêt ou énumération des « liens » avec d'autres éléments de la population), un échantillon de suivi s_1 est sélectionné parmi la population restante, en utilisant un plan de sondage conditionnel $p_1(s_1 | s_0)$. Ce processus est répété pour des échantillons additionnels successifs s_2, s_3, \dots jusqu'à ce que soit satisfait un critère cible, tel que la taille totale de l'échantillon ou le nombre de « vagues » d'échantillonnage. L'échantillon final correspond à l'union de tous les échantillons successifs. La façon détaillée dont les échantillons successifs sont tirés varie selon le plan d'échantillonnage adaptatif. La section 2.2 de l'article de Thompson publié dans le présent numéro et Thompson (2006) contiennent d'autres renseignements sur l'échantillonnage en ligne adaptatif, un type très souple d'échantillonnage adaptatif qui englobe un grand nombre des autres plans de sondage en tant que cas particuliers.

Comme les plans d'échantillonnage pour chacune des vagues d'échantillonnage sont des plans probabilistes, il est possible d'obtenir des estimateurs valides sous le plan. Un estimateur simple de la moyenne de population finie $\mu_N = N^{-1} \sum_{i=1}^N y_i$ est construit de la façon suivante. Partant du plan de sondage initial p_0 avec les probabilités d'inclusion connexes π_{0i} , un estimateur sans biais de la moyenne de population est donné par $\hat{\mu}_0 = N^{-1} \sum_{i=1}^N y_i / \pi_{0i}$. Pour chacune des vagues subséquentes d'échantillonnage $k = 1, \dots, K$,

- Kiaer, A. (1899). Sur les méthodes représentatives ou typologiques appliquées à la statistique. *Bulletin de l'Institut International de Statistique*, 11, 1, 180-185.
- Kiaer, A. (1903). Sur les méthodes représentatives ou typologiques appliquées à la statistique. *Bulletin de l'Institut International de Statistique*, 13, 1, 66-78.
- Kiaer, A. (1905). Discours sans intitulé sur la méthode représentative. *Bulletin de l'Institut International de Statistique*, 14, 1, 119-134.
- Kott, P.S. (1986). When a mean-of-ratios is the best linear unbiased estimator under a model. *The American Statistician*, 40, 202-204.
- Langel, M., et Tillé, Y. (2010). Corrado Gini, a pioneer in balanced sampling and inequality theory. Rapport technique, University of Neuchâtel.
- Legg, J.C., et Yu, C.T. (2010). Comparaison de méthodes de restriction de l'ensemble d'échantillons. *Techniques d'enquête*, 36, 75-87.
- Lesage, E. (2008). Contraintes d'équilibrage non linéaires. Dans *Méthodes d'enquêtes : applications aux enquêtes longitudinales, à la santé et aux enquêtes électrolales*, (Eds., P. Guilbert, D. Haziza, A. Ruiz-Gazen et Y. Tillé), Paris. Dunod, 285-289.
- Mart, G., Barbarà, G., Mitias, G. et Passamonti, S. (2007a). Construction de un estimador de variancia para muestras balanceadas estratificadas. Dans *XXV Coloquio Argentino de Estadística. Mar del Plata, L'Argentine. 22, 23 y 24 de Octubre de 2007*.
- Mart, G., Barbarà, G., Mitias, G. et Passamonti, S. (2007b). Muestras equilibradas en poblaciones finitas: un estudio comparativo en muestras de explotaciones agropecuarias. Dans *Undécimas Jornadas "Investigaciones en la Facultad" de Ciencias Económicas y Estadística, noviembre de 2007*, Universidad Nacional de Rosario, L'Argentine.
- Nedyalkova, D., Péa, J. et Tillé, Y. (2006). A review of some current methods of coordination of stratified samples. introduction and comparison of new methods based on microstrata. Rapport technique, Université de Neuchâtel.
- Nedyalkova, D., and Tillé, Y. (2009). Optimal sampling and estimation strategies under linear model. *Biometrika*, 95, 521-537.
- Nedyalkova, D., and Tillé, Y. (2010). Bias robustness and efficiency in model-based inference. Rapport technique, Université de Neuchâtel.
- Neyman, J. (1934). On the two different aspects of representative purposive selection. *Journal of the Royal Statistical Society*, 97, 558-606.
- Neyman, J. (1952). *Lectures and Conferences on Mathematical Statistics and Probability*. Graduate School, U.S. Department of Agriculture, Washington.
- Pétié, P. (2008). Échantillonnage à entropie maximale sous contraintes : un algorithme rapide basé sur l'optimisation linéaire en nombres binaires. In *Méthodes d'enquêtes : applications aux enquêtes longitudinales, à la santé et aux enquêtes électrolales*, (Eds., P. Guilbert, D. Haziza, A. Ruiz-Gazen et Y. Tillé), Paris. Dunod, 294-299.
- Rivière, P. (1999). Coordination of samples: The microstrata methodology. Dans *13th International Roundtable on Business Survey Frames*, Paris. Insee.
- Kiaer, A. (1903). Sur les méthodes représentatives ou typologiques appliquées à la statistique. *Bulletin de l'Institut International de Statistique*, 11, 1, 180-185.
- Kiaer, A. (1905). Discours sans intitulé sur la méthode représentative. *Bulletin de l'Institut International de Statistique*, 14, 1, 119-134.
- Kott, P.S. (1986). When a mean-of-ratios is the best linear unbiased estimator under a model. *The American Statistician*, 40, 202-204.
- Langel, M., et Tillé, Y. (2010). Corrado Gini, a pioneer in balanced sampling and inequality theory. Rapport technique, University of Neuchâtel.
- Legg, J.C., et Yu, C.T. (2010). Comparaison de méthodes de restriction de l'ensemble d'échantillons. *Techniques d'enquête*, 36, 75-87.
- Lesage, E. (2008). Contraintes d'équilibrage non linéaires. Dans *Méthodes d'enquêtes : applications aux enquêtes longitudinales, à la santé et aux enquêtes électrolales*, (Eds., P. Guilbert, D. Haziza, A. Ruiz-Gazen et Y. Tillé), Paris. Dunod, 285-289.
- Mart, G., Barbarà, G., Mitias, G. et Passamonti, S. (2007a). Construction de un estimador de variancia para muestras balanceadas estratificadas. Dans *XXV Coloquio Argentino de Estadística. Mar del Plata, L'Argentine. 22, 23 y 24 de Octubre de 2007*.
- Mart, G., Barbarà, G., Mitias, G. et Passamonti, S. (2007b). Muestras equilibradas en poblaciones finitas: un estudio comparativo en muestras de explotaciones agropecuarias. Dans *Undécimas Jornadas "Investigaciones en la Facultad" de Ciencias Económicas y Estadística, noviembre de 2007*, Universidad Nacional de Rosario, L'Argentine.
- Nedyalkova, D., Péa, J. et Tillé, Y. (2006). A review of some current methods of coordination of stratified samples. introduction and comparison of new methods based on microstrata. Rapport technique, Université de Neuchâtel.
- Nedyalkova, D., and Tillé, Y. (2009). Optimal sampling and estimation strategies under linear model. *Biometrika*, 95, 521-537.
- Nedyalkova, D., and Tillé, Y. (2010). Bias robustness and efficiency in model-based inference. Rapport technique, Université de Neuchâtel.
- Neyman, J. (1934). On the two different aspects of representative purposive selection. *Journal of the Royal Statistical Society*, 97, 558-606.
- Neyman, J. (1952). *Lectures and Conferences on Mathematical Statistics and Probability*. Graduate School, U.S. Department of Agriculture, Washington.
- Pétié, P. (2008). Échantillonnage à entropie maximale sous contraintes : un algorithme rapide basé sur l'optimisation linéaire en nombres binaires. In *Méthodes d'enquêtes : applications aux enquêtes longitudinales, à la santé et aux enquêtes électrolales*, (Eds., P. Guilbert, D. Haziza, A. Ruiz-Gazen et Y. Tillé), Paris. Dunod, 294-299.
- Rivière, P. (1999). Coordination of samples: The microstrata methodology. Dans *13th International Roundtable on Business Survey Frames*, Paris. Insee.
- Statistique Canada, N° 12-001-X au catalogue
- Yates, F. (1946). A review of recent statistical developments in sampling and sampling surveys. *Journal of the Royal Statistical Society*, A109, 12-43.
- Yates, F. (1960). *Sampling Methods for Censuses and Surveys*. Charles Griffin, Londres, L'Angleterre, troisième édition.
- Roussseau, S., et Tardieu, F. (2004). La macro SAS CUBE d'échantillonnage équilibré. Documentation de l'utilisateur. Rapport technique, Insee, Paris.
- Royall, R.M. (1976a). Likelihood functions in finite population sampling theory. *Biometrika*, 63, 605-614.
- Royall, R.M. (1976b). The linear least squares prediction approach to two-stage sampling. *Journal of the American Statistical Association*, 71, 657-664.
- Royall, R.M. (1988). The prediction approach to sampling theory. In *Handbook of Statistics Volume 6: Sampling*, (Eds., P.R. Krishnaiah et C.R. Rao), Amsterdam. Elsevier/North-Holland, 399-413.
- Royall, R.M., et Pfeffermann, D. (1982). Balanced samples and robust bayesian inference in finite population sampling. *Biometrika*, 69, 401-409.
- Sunter, A. (1977). List sequential sampling with equal or unequal probabilities without replacement. *Applied Statistics*, 26, 261-268.
- Tardieu, F. (2001). Échantillonnage équilibré: de la théorie à la pratique. Rapport technique, Insee, Paris.
- Thionet, P. (1953). *La théorie des sondages*. Insee, Imprimerie nationale, Paris.
- Tillé, Y. (2001). *Théorie des sondages : échantillonnage et estimation en populations finies*. Dunod, Paris.
- Tillé, Y. (2006a). Balanced sampling by means of the cube method. Dans *Joint Statistical Meeting of the American Statistical Association*, Seattle août 2006.
- Tillé, Y. (2006b). *Sampling Algorithms*. New York : Springer.
- Tillé, Y., et Favre, A.-C. (2004). Co-ordination, combination and extension of optimal balanced samples. *Biometrika*, 91, 913-927.
- Tillé, Y., et Favre, A.-C. (2005). Optimal allocation in balanced sampling. *Statistics and Probability Letters*, 74, 31-37.
- Tillé, Y., et Matei, A. (2007). *The R Package Sampling*. The Comprehensive R Archive Network, <http://cran.r-project.org/>, Manual of the Contributed Packages.
- Titani, M. (2006). Le plan de sondage équilibré et l'estimation du total d'une population finie. Dans *Méthodes d'enquêtes et sondages : pratiques européennes et nord-américaines*, (Eds., P. Lavallée et L.-P. Rivest), Paris, Dunod, 411-416.
- Valliant, R., Dorfman, A.H. et Royall, R.M. (2000). *Finite Population Sampling and Inference: A Prediction Approach*. New York : John Wiley & Sons, Inc.
- Wilms, L. (2000). Présentation de l'échantillon-maître en 1999 et application au tirage des unités primaires par la macro cube. Dans *Séries Insee Méthodes : Actes des Journées de Méthodologie Statistique*, Paris. Insee.
- Yates, F. (1946). A review of recent statistical developments in sampling and sampling surveys. *Journal of the Royal Statistical Society*, A109, 12-43.

- De Ree, S.J.M. (1999). Co-ordination of business samples using measured response burden. Dans *Contributed paper, 52th Session of the ISI Helsinki*.
- Desplanques, G. (2000). La rénovation du recensement de la population. Dans *Actes de la séance du 5 octobre 2000 du séminaire méthodologique SFD5-Insee sur la rénovation du recensement*, 2-5.
- Dessertine, A. (2006). Sondages et séries temporelles : une application pour la prévision de la consommation électrique. Dans *Actes des journées Françaises de Statistique 2006*, Clamart, France.
- Dessertine, A. (2007). Sampling and data-stream: Some ideas to built balanced sampling using auxiliary Hilbertian informations. Dans *Proceedings of 56th the International Statistical Institute Conference: IPM56 - New methods of sampling*, Lisboa, Portugal.
- Deville, J.-C. (1992). Constrained samples, conditional inference, weighting: Three aspects of the utilisation of auxiliary information. Dans *Proceedings of the Workshop on the Uses of Auxiliary Information in Surveys*, Örebro (La Suède).
- Deville, J.-C. (1998). La correction de la non-réponse par calage ou par échantillonnage équilibré. Dans *Recueil de la Section des méthodes d'enquêtes des communications présentées au 26^{ème} congrès de la Société Statistique du Canada*, 103-110, Sherbrooke.
- Deville, J.-C. (2006). Stochastic imputation using balanced sampling. Dans *Joint Statistical Meeting of the American Statistical Association*, Seattle août 2006.
- Deville, J.-C., Grosbras, J.-M. et Roth, N. (1988). Efficient sampling algorithms and balanced sample. Dans *COMPSTAT, Proceedings in Computational Statistics*, Heidelberg. Physica Verlag, 255-266.
- Deville, J.-C., et Tillé, Y. (1998). Unequal probability sampling without replacement through a splitting method. *Biometrika*, 85, 89-101.
- Deville, J.-C., et Tillé, Y. (2004). Efficient balanced sampling: The cube method. *Biometrika*, 91, 893-912.
- Deville, J.-C., et Tillé, Y. (2005). Variance approximation under balanced sampling. *Journal of Statistical Planning and Inference*, 128, 569-591.
- Dudoignon, L., et Vanheuverzwyn, A. (2006). Tirage d'un échantillon à probabilités inégales : application au panel Médiamat. Dans *Actes de des Journées de Méthodologie Statistique*, 1-10.
- Dumais, J., Bertrand, P. et Kauffmann, B. (2000). Sondage, estimation et précision dans la rénovation du recensement de la population. Dans *Actes de la séance du 5 octobre 2000 du séminaire méthodologique SFD5-Insee sur la rénovation du recensement*, 6-26.
- Dumais, J., et Isnard, M. (2000). Le sondage de logements dans les grandes communes dans le cadre du recensement renoué de la population. Dans *Séries Insee Méthodes : Actes des Journées de Méthodologie Statistique*, Paris. Insee, 100, 37-76.
- Durt, J.-M., et Dumais, J. (2001). La rénovation du recensement Français. Dans *le Recueil : Symposium 2001, La Qualité des Données d'un Organisme Statistique : Une Perspective Méthodologique*, Statistique Canada, Ottawa.
- Durt, J.-M., et Dumais, J. (2002). La rénovation du recensement Français. *Techniques d'enquête*, 28, 47-53.
- Even, K. (2002). Improved tool for evaluating employment and vocational training policy: Panel of beneficiaries. *Premières Informations Synthèses, Direction de l'Animation de la Recherche des Etudes et des Statistiques (DARES)* du Ministère du Travail des relations sociales et de la solidarité, 33, 1, 1-7.
- Falorsi, P.D., et Righi, P. (2008). Une approche d'échantillonnage équilibré pour des plans de sondage à stratification multidimensionnelle pour l'estimation pour petits domaines. *Techniques d'enquête*, 34, 247-259.
- Fecleau, S., et Jocelyn, W. (2006). Une application de l'échantillonnage équilibré : le plan de sondage des entreprises non incorporées. Dans *Méthodes d'enquêtes et sondages : pratiques européennes et nord-américaines*, (Eds., P. Lavallée et L.-P. Rivest), Paris. Dunod, 405-410.
- Fuller, W.A. (2009). Some design properties of a rejective sampling procedure. *Biometrika*, 96, 933-944.
- Fuller, W.A. (2010). Replication variance estimation for rejective sampling. Dans *Seminar of Statistics Canada*, June 2010, Ottawa.
- Gini, C. (1928). Une application de la méthode représentative aux matériaux du dernier recensement de la population italienne (1^{er} décembre 1921). *Bulletin of the International Statistical Institute*, 23, 2, 198-215.
- Gini, C., et Galvani, L. (1929). Di una applicazione del metodo rappresentativo all'ultimo censimento Italiano della popolazione (1^o decembre, 1921). *Annali di Statistica*, Série 6, 4, 1-107.
- Gismondi, R. (2007). Quick estimation of tourist nights spent in Italy. *Statistical Methods and Applications*, 16, 141-168.
- Häjek, J. (1981). *Sampling from a Finite Population*. New York : Marcel Dekker.
- Hedayat, A.S., et Majumdar, D. (1995). Generating desirable sampling plans by the technique of trade-off in experimental design. *Journal of Statistical Planning and Inference*, 44, 237-247.
- Hesse, C. (1998). Sampling co-ordination: A review by country. Rapport technique E9908, Direction des Statistique d'Entreprises, Insee, Paris.
- Jocelyn, W. (2006). Sampling and estimation strategies for the canadian unincorporated business population. Dans *Joint Statistical Meeting of the American Statistical Association*, Seattle août 2006.
- Kjaer, A. (1896). Observations et expériences concernant des dénombremments représentatifs. *Bulletin de l'Institut International de Statistique*, 9, 2, 176-183.
- Statistique Canada, N° 12-001-X au catalogue

application comptant 40 variables équilibrées. Afin de sélectionner l'échantillon, les temps de calcul augmentent proportionnellement à $N \times p^2$, où N est la taille de population et p est le nombre de variables d'équilibrage. Il est donc possible de tirer un échantillon dans une population de plusieurs millions d'unités statistiques.

Remerciements

Le présent article a été rédigé à la suite d'une invitation à présenter une communication à la conférence de la Demographic Statistical Methods Division du U.S. Census Bureau, tenue en juin 2008. L'auteur remercie le U.S. Census Bureau, en particulier Patrick Flanagan, sans lequel le présent article n'aurait jamais été rédigé. L'auteur remercie également un rédacteur associé et deux examinateurs anonymes de leurs commentaires et corrections fort utiles qui l'ont aidé à améliorer le manuscrit.

Bibliographie

Chauvet, G. (2007). *Méthodes de Bootstrap en Population Finie*. Thèse de doctorat, Université Rennes 2.

Chauvet, G. (2009). Échantillonnage équilibré stratifié. *Techniques d'enquête*, 35, 123-127.

Chauvet, G., Bonnerly, D. et Deville, J.-C. (2010a). Optimal inclusion probabilities for balanced sampling. *Journal of Statistical Planning and Inference*, 141, 2, 984-994.

Chauvet, G., Deville, J. et Haziza, D. (2010b). Adapting the cube algorithm for balanced random imputation in surveys. Rapport technique, Ensaï, Rennes.

Chauvet, G., Deville, J. et Haziza, D. (2011). On balanced random imputation in surveys. *Biometrika*.

Chauvet, G., et Tillé, Y. (2005a). *Fast SAS Macros for balancing Samples: user's guide*. Software Manual, Université de Neuchâtel, <http://www2.unine.ch/statistics/page10890.html>.

Chauvet, G., et Tillé, Y. (2005b). New SAS macros for balanced sampling. Dans *Journées de Méthodologie Statistique*, Insee, Paris.

Chauvet, G., et Tillé, Y. (2006). A fast algorithm of balanced sampling. *Journal of Computational Statistics*, 21, 9-31.

Chipperfield, J. (2009). An evaluation of cube sampling for ABS household surveys. Rapport technique, Australian Bureau of Statistics.

Christine, M. (2006). Use of balanced sampling in the framework of the master sample for french household surveys. Dans *Joint Statistical Meeting of the American Statistical Association*, Seattle août 2006.

Christine, M., et Wilms, L. (2003). Problèmes théoriques et pratiques de la construction de l'« EMEX » : comment améliorer la précision des extensions régionales des enquêtes nationales grâce à un échantillonnage additionnel ? Dans *le Recueil : Symposium 2003, Défis Reliés à la Réalisation d'Enquêtes pour la Prochaine Décennie*, Statistique Canada, Ottawa.

Cumberland, W.G., et Royall, R.M. (1981). Prediction models in unequal probability sampling. *Journal of the Royal Statistical Society*, B, 43, 353-367.

Cumberland, W.G., et Royall, R.M. (1988). Does simple random sampling provide adequate balance? *Journal of the Royal Statistical Society*, B, 50, 118-124.

da Silva, A.D., da Silva Borges, A., Aires Leme, R. et Moura Reis Miceil, A.P. (2006). Modalidades alternativas de censo demográfico: o cenário internacional a partir das experiências dos estados unidos, França, Holanda, Israël e Alemanha. Rapport technique, Instituto Brasileiro de Geografia e Estatística.

D'Alò, M., Di Consiglio, L., Falorsi, S. et Solari, F. (2006). Small area estimation of the Italian poverty rate. *Statistics in Transition*, 7, 771-784.

plusieurs millions d'unités statistiques.

Remerciements

Le présent article a été rédigé à la suite d'une invitation à présenter une communication à la conférence de la Demographic Statistical Methods Division du U.S. Census Bureau, tenue en juin 2008. L'auteur remercie le U.S. Census Bureau, en particulier Patrick Flanagan, sans lequel le présent article n'aurait jamais été rédigé. L'auteur remercie également un rédacteur associé et deux examinateurs anonymes de leurs commentaires et corrections fort utiles qui l'ont aidé à améliorer le manuscrit.

Bibliographie

Ardilly, P. (1991). Échantillonnage représentatif optimum à probabilités inégales. *Annales d'Economie et de Statistique*, 23, 91-113.

Ardilly, P. (2006). *Les Techniques de Sondage*. Technip, Paris.

Bardaji, J. (2001). Un an après la sortie d'un contrat emploi consolidé : près de six chances sur dix d'avoir un emploi. *Premières Informations Synthèses, Direction de l'Animation de la Recherche des Etudes et des Statistiques (DARES) du Ministère du Travail des relations sociales et de la solidarité*, 43, 3, 1-8.

Bertrand, P., Christian, B., Chauvet, G. et Grosbras, J.-M. (2004). Plans de sondage pour le recensement renové de la population. Dans *Séries Insee Méthodes : Actes des Journées de Méthodologie Statistique*, Paris. Insee.

Biggeri, L., et Falorsi, P.D. (2006). A probability sample strategy for improving the quality of the consumer price index survey using the information of the business register. Dans *Proceedings of the Conference of European Statisticians Group of Experts on Consumer Price Indices*, huitième réunion, Genève, 10-12 mai 2006.

Bousabaa, A., Lieber, J. et Strolli, R. (1999). La macro cube. Rapport technique, Ensaï, Rennes.

Breidt, F.J., et Chauvet, G. (2010a). Improved variance estimation for balanced samples drawn via the cube method. *Journal of Statistical Planning and Inference*, 141, 479-487.

Breidt, F.J., et Chauvet, G. (2010b). Penalized balanced sampling. Document de travail, Ensaï.

Brewer, K.R.W. (1975). A simple procedure for rpswor. *Australian Journal of Statistics*, 17, 166-172.

où $O_p(x)/x$ est une quantité qui demeure bornée en probabilité quand x tend vers l'infini.

Les gains de précision sont par conséquent considérables.

Le petit problème d'arrondi peut être résolu par un petit calage. Le problème d'arrondi est dû au fait que le tirage d'un échantillon est un problème en nombres entiers. Il se produit également dans la stratification, qui est un cas particulier de l'équilibrage. Dans le cas de la stratification avec répartition proportionnelle, les sommes des probabilités d'inclusion dans les strates ne sont généralement pas des entiers. Donc, les tailles d'échantillon de strate sont obtenues en arrondissant la somme des probabilités d'inclusion dans les strates. Dans la méthode du cube, cet arrondi est effectué automatiquement et aléatoirement de manière à s'assurer que les probabilités d'inclusion soient exactement satisfaites.

7.6 Échantillonnage équilibré dans les enquêtes répétées

L'échantillonnage répété pose un problème important. Celui-ci tient au fait que, si un échantillon équilibré est obtenu par tirage à probabilités d'inclusion inégales, l'échantillon complémentaire n'est pas nécessairement équilibré. En effet, l'égalité

$$\sum_{k \in S} \frac{\pi_k}{x_k} = \sum_{k \in U} \pi_k$$

n'implique pas que

$$\sum_{k \in U \setminus S} \frac{1 - \pi_k}{x_k} = \sum_{k \in U} \pi_k.$$

Ce problème s'est produit dans l'échantillon-maître français. Dans ce plan de sondage, les unités primaires, qui sont des secteurs géographiques, sont sélectionnées avec probabilités inégales qui sont proportionnelles à la taille. Après le tirage de l'échantillon, certaines régions ont demandé des échantillons complémentaires de secteurs qui n'avaient pas été sélectionnés. Cette question est complexe, parce que l'échantillon complémentaire d'un échantillon équilibré n'est plus équilibré et que le but est donc de tirer un échantillon équilibré dans une partie de la population qui n'est plus équilibrée. Tillé et Favre (2004) ont donné quelques méthodes pour coordonner des échantillons équilibrés qui ont été sélectionnées avec probabilités d'inclusion inégales. De manière plus générale, la coordination (au sens de gestion du chevauchement) des échantillons équilibrés peut être difficile quand le plan de sondage est équilibré. Bien que cela ne soit pas facile, il est possible d'organiser des rotations si tous les échantillons sont sélectionnés ensemble selon un tirage à probabilités d'inclusion égales. En effet, dans ce cas, le complètement $S = U \setminus S$ des échantillons S est également un échantillon équilibré. Un

7.7 Principales implémentations de l'échantillonnage équilibré

Une application SAS/IML[®] a d'abord été programmée par trois étudiants de l'École Nationale de la Statistique et de l'Analyse de l'Information (ENSAI) (Bousabaa et coll. 1999). La version officielle de l'Institut national de la statistique et des études économiques (Insee), produite par Tardieu (2001) et par Rousseau et Tardieu (2004), est maintenant accessible sur le site Web de l'Insee. Une autre version SAS/IML[®] produite par Chauvet et Tillé (2005a, 2005b, 2006) est également disponible sur le site Web de l'Université de Neuchâtel. En langage R, le logiciel d'échantillonnage (Tillé et Matei 2007) permet d'utiliser la méthode du cube. Ces logiciels sont gratuits, accessibles sur Internet et faciles à utiliser.

Les programmes existants écrits au moyen du langage R ou SAS/IML[®] ne présentent aucune limite en ce qui concerne la taille de population. Il est possible d'exécuter une

$$\sum_{k \in S} \frac{x_k / (2\pi_k)}{1/2} = \sum_{k \in S} \frac{x_k}{2\pi_k} = \sum_{k \in U} x_k.$$

Ensuite, il faut sélectionner un échantillon S_1 à partir de S_1 . Cet échantillon doit être sélectionné avec la probabilité d'inclusion $\pi_{k1} = 0,5$ et doit être équilibré sur $x_k / 2\pi_k$, ce qui donne les équations d'équilibrage suivantes :

$$\sum_{k \in S} \frac{x_k}{2\pi_k} = \sum_{k \in U} x_k.$$

telles que

Si les échantillons sont sélectionnés avec probabilités d'inclusion inégales, certaines solutions sont décrites dans Tillé et Favre (2004). Un cas particulier intéressant peut être résolu facilement, c'est-à-dire quand deux échantillons non chevauchants doivent être sélectionnés avec les mêmes probabilités d'inclusion inégales $\pi_k < 0,5$ pour la même population. D'abord, il faut sélectionner un échantillon S_1 équilibré sur x_k avec les probabilités d'inclusion $\pi_{k1} = 2\pi_k$ telles que

Les cinq groupes correspondent à cinq échantillons équilibrés de communes.

deuxième échantillon équilibré peut être tiré directement de S et ainsi de suite. Cette méthode a été utilisée pour créer cinq groupes de rotation dans l'échantillon-maître français.

variables auxiliaires sont fausses. Le gain d'efficacité dépend uniquement de la corrélation entre les variables d'équilibrage et les variables d'intérêt. Cette corrélation est rarement affectée par les erreurs touchant les variables d'équilibrage.

Plusieurs variables peuvent être utilisées pour améliorer les estimations sur petits domaines. Afin de s'assurer qu'un domaine D n'est pas vide, on peut simplement ajouter la variable auxiliaire :

$$x_k = \begin{cases} \pi_k & \text{si } k \in D \\ 0 & \text{autrement,} \end{cases}$$

qui implique que le nombre d'unités échantillonnées qui appartiennent à D est égal à

$$n_D = \sum_{k \in D} x_k = \sum_{k \in D} \pi_k,$$

si n_D est un entier, ou à l'un des deux entiers les plus proches de n_D si n_D n'est pas un entier.

Dans certains cas, il est utile d'équilibrer sur des variables auxiliaires dans des sous-groupes, des domaines ou des strates. Une procédure intéressante décrite dans Chauvet (2009) consiste à exécuter séparément la phase de vol dans chaque strate. Un problème d'arrondi surviendra alors dans chaque strate. Il est possible de fusionner ces problèmes d'arrondi et d'exécuter de nouveau une phase de vol sur l'ensemble de la population. Enfin, la phase d'atterrissage est appliquée uniquement à l'ensemble de la population. Cette procédure permet que les équations d'équilibrage soient approximativement satisfaites dans chaque strate sans cumuler les problèmes d'arrondi.

Les probabilités d'inclusion doivent être calculées avant l'échantillonnage. Si l'on émet l'hypothèse d'un modèle linéaire, ces probabilités doivent, en principe, être proportionnelles aux erreurs du modèle afin de minimiser la variance (voir Tillé et Favre 2005 ; Chauvet, Bonney et Deville 2010a ; Nedyalkova et Tillé 2009, 2010). Ce choix généralise la méthode d'allocation de Neyman pour l'échantillonnage stratifié (Neyman 1934). Cependant, les probabilités d'inclusion doivent souvent être choisies en fonction d'autres contraintes. Par exemple, afin de construire les groupes de rotation du recensement français, les probabilités d'inclusion doivent être égales à un cinquième.

7.4 Équilibrage versus calage

La stratification est un cas particulier de l'équilibrage, tandis que la poststratification est un cas particulier du calage. Dans le cas de la stratification et de l'équilibrage, les pondérations ne deviennent pas aléatoires. Donc, il s'agit généralement d'une meilleure stratégie. Néanmoins, l'équilibrage requiert une plus grande quantité d'information auxiliaire. En effet, dans l'échantillonnage équilibré, les

En général, il est recommandé d'effectuer de nouveau le calage sur les variables d'équilibrage à l'étape de l'estimation, même si un plus grand nombre de variables de calage sont disponibles. Si l'on n'utilise que de nouvelles variables pour le calage, l'effet de l'équilibrage risque d'être perdu. Il existe toutefois un cas où le calage peut être utilisé sans effectuer un nouveau calage sur les variables d'équilibrage. Il s'agit de la situation où, conditionnellement aux variables de calage, nous pouvons raisonnablement supposer que les variables d'équilibrage ne sont plus corrélées aux variables d'intérêt. Cela peut se produire quand les variables d'équilibrage et de calage sont les mêmes variables mesurées à des périodes différentes et que les variables de calage sont plus récentes.

Quand le coefficient de détermination entre la variable d'intérêt et les variables auxiliaires est égal à un ou proche de cette valeur, le calage est plus efficace en raison du problème d'arrondi de l'échantillonnage équilibré. Quoiqu'il en soit, la meilleure stratégie consiste toujours à utiliser de concert l'échantillonnage équilibré et le calage (voir la simulation dans Deville et Tillé 2004).

7.5 Précision des équations d'équilibrage

Il est possible de prouver, sous des hypothèses raisonnables (voir Deville et Tillé 2004), qu'avec la méthode du cube,

$$\left| \frac{\widehat{X}_j}{X_j} - \frac{X_j}{X_j} \right| > O(p/n),$$

où p est le nombre de variables et $O(x)/x$ est une quantité qui reste bornée quand x tend vers l'infini. Sous échantillonnage aléatoire simple,

$$\left| \frac{X_j}{X_j - X_j} \right| = O_p(\sqrt{1/n}),$$

À Statistique Canada, Fecteau et Jocelyn (2006) et Jocelyn (2006) ont testé l'échantillonnage équilibré pour tirer un échantillon d'entreprises. Les entreprises canadiennes non constituées en société produisent leur déclaration de revenus sur papier ou électroniquement. Plus de la moitié des déclarations sont soumises électroniquement. L'échantillonnage équilibré a été utilisé pour sélectionner un échantillon parmi les entreprises ayant produit une déclaration électronique, de manière que, pour certaines variables clés dont la valeur est connue pour l'ensemble de la population, les moyennes d'échantillon concordent avec les moyennes de population connues.

L'échantillonnage équilibré peut également être utilisé pour imputer une valeur manquante en cas de non-réponse partielle. En effet, l'utilisation d'un modèle pour prédire une imputation attribue les valeurs centrales, ce qui donne lieu à une inférence biaisée sur les quantiles. Par contre, une imputation aléatoire augmente généralement les variances des estimateurs. Afin de résoudre ce dilemme, Deville (1998, 2005, 2006), ainsi que Chauvet, Deville et Haziza (2010b, 2010c) ont proposé d'utiliser l'imputation par prédiction et d'ajouter un résidu qui est choisi parmi les résidus des répondants selon un plan de sondage équilibré. Ce faisant, on évite d'ajouter un terme de variance au total de la variable imputée.

7.2 Échantillonnage équilibré versus d'autres techniques d'échantillonnage

L'échantillonnage à probabilités inégales est un cas particulier de la méthode du cube. En effet, quand la probabilité d'inclusion est la seule variable auxiliaire, la taille de l'échantillon est fixe. La méthode du cube est une généralisation de la méthode de scission (voir Deville et Tillé 1998), qui comporte plusieurs algorithmes d'échantillonnage à probabilités inégales (méthode de Brewer, méthode du pivot, méthode corrigée de Sunter, voir Brewer 1975 ; Sunter 1977 ; Deville et Tillé 1998 ; Tillé 2006b). La stratification est aussi un cas particulier d'échantillonnage équilibré. La méthode du cube permet d'effectuer l'échantillonnage sur des strates chevauchantes et d'utiliser ensemble des variables qualitatives et quantitatives. Même l'échantillonnage systématique peut être vu comme un échantillonnage équilibré sur la statistique d'ordre reliée à la variable en fonction de laquelle la population est ordonnée.

Presque toutes les autres techniques d'échantillonnage sont des cas particuliers d'échantillonnage équilibré (sauf l'échantillonnage à plusieurs degrés). En fait, l'échantillonnage équilibré est simplement plus général, en ce sens que toutes les autres méthodes d'échantillonnage peuvent être mises en œuvre en se servant de la méthode du cube. Cette dernière permet d'utiliser n'importe quelle variable

7.3 Choix de la stratégie d'échantillonnage

Il est bien connu que l'estimateur par le ratio et l'estimateur poststratifié sont des cas particuliers de l'estimateur par la régression. Ce dernier est aussi un cas particulier de l'estimateur par calage (qui comprend un ajustement non linéaire). De même, l'échantillonnage équilibré est une méthode d'échantillonnage plus générale qui englobe presque toutes les autres. L'algorithme de la méthode du cube peut paraître compliqué, mais, une fois implémenté, il permet d'exécuter une fonction avec deux arguments, à savoir le vecteur des probabilités d'inclusion et la matrice des variables d'échantillonnage.

La principale recommandation est de choisir des variables d'échantillonnage qui sont étroitement corrélées aux variables d'intérêt. Comme dans tout problème de régression, les variables d'échantillonnage doivent être choisies parcimonieusement : il ne faut pas en choisir un trop grand nombre, parce que la précision n'augmente plus une fois que le nombre de variables est grand et l'instabilité de l'estimateur de variance s'accroît avec chaque variable supplémentaire. En pratique, le but n'est pas d'estimer une variable, mais un ensemble de variables d'intérêt. Donc, l'ensemble de variables auxiliaires doit être corrélé à toutes les variables d'intérêt. De surcroît, les variables auxiliaires ne doivent pas être trop corrélées entre elles.

Lesage (2008) a proposé une méthode pour équilibrer un échantillon sur des statistiques complexes au lieu d'utiliser simplement les totaux de population. L'idée fondamentale consiste à effectuer l'échantillonnage sur la valeur linéarisée (ou fonction d'influence) du paramètre d'intérêt. Breidt et Chauvet (2010b) ont proposé de recourir à l'échantillonnage équilibré pénalisé afin de pouvoir éventuellement relâcher certaines contraintes d'échantillonnage, ce qui peut être utile par exemple dans l'estimation sur petits domaines.

Dans de nombreux cas, les variables d'échantillonnage contiennent des erreurs de mesure. Ainsi, dans la plupart des registres, on peut soupçonner la présence d'erreurs dans les données. Des valeurs manquantes peuvent manifestement exister et les variables auxiliaires sont souvent corrigées par une méthode d'imputation. Pour ce qui est du calage, le fait que les variables auxiliaires contiennent des erreurs n'est pas très important pourvu que le calage soit effectué sur le total des variables auxiliaires du registre. En effet, sous échantillonnage équilibré, on utilise l'estimateur de Horvitz-Thompson qui est sans biais même si les variables

équilibré, les variances des estimateurs de Horvitz-Thompson de la variable d'intérêt dépendra uniquement des résidus du modèle.

Les avantages de l'échantillonnage équilibré sont les suivants :

- i) L'échantillonnage équilibré augmente la précision de l'estimateur de Horvitz-Thompson. Ce point a été traité à la section 6. En effet, la variance de l'estimateur de Horvitz-Thompson dépend uniquement des résidus de la régression de la variable d'intérêt en fonction des variables d'équilibrage.
- ii) L'échantillonnage équilibré protège contre les grandes erreurs d'échantillonnage. En effet, les échantillons les moins favorables ont une probabilité nulle d'être tirés.
- iii) Si la variable d'intérêt est bien expliquée par l'information auxiliaire, dans l'inférence basée sur un modèle, l'échantillonnage équilibré protège contre une erreur de spécification du modèle. Ce point est traité en détail par Royall (1976a, 1976b) et par Valliant et coll. (2000). Une discussion récente de cette question importante est présentée dans Nedelkova et Tillé (2009, 2010).
- iv) L'échantillonnage équilibré permet de s'assurer que les tailles d'échantillon dans les domaines prévus ne soient pas trop faibles ou – situation pire encore – nulles. En effet, si une variable indicatrice du domaine est ajoutée à la liste des variables auxiliaires, la taille du domaine est alors fixée dans l'échantillon.
- v) L'échantillonnage équilibré permet d'éviter les pondérations aléatoires. Sous échantillonnage équilibré, nous pouvons utiliser les pondérations de Horvitz-Thompson. Si le plan de sondage ne contient aucune contrainte d'équilibrage (par exemple sous échantillonnage poissonnien), le système de pondération obtenu par une procédure de calage devient très aléatoire, ce qui augmente la variance des estimateurs. Si l'échantillon est équilibré, les pondérations seront moins aléatoires, même si une procédure de calage est utilisée après l'équilibrage.

L'existence de propriétés faciles à utiliser a contribué à l'usage répandu de la méthode du cube dans plusieurs processus statistiques importants. La première grande application de la méthode du cube est la sélection de groupes de rotation pour le recensement français (voir Desplanques 2000 ; Dumas, Bertrand et Kauffmann 2000 ; Durt et Dumas 2001, 2002 ; Dumas et Isnard 2000 ;

Bertrand, Christian, Chauvet et Grosbras 2004 ; da Silva, da Silva Borges, Aires Leme et Moura Reis Miceli 2006). Pour les communes de moins de 10 000 habitants, cinq groupes de rotation non chevauchants de communes sont sélectionnés selon un plan de sondage équilibré avec probabilités d'inclusion de 8 %. Donc, après cinq ans, une visite est effectuée à 40 % des adresses. Les variables d'équilibrage sont des variables sociodémographiques tirées du dernier recensement.

Dans l'échantillon-mère français, les unités primaires sont les régions géographiques qui sont sélectionnées selon un plan de sondage équilibré (voir Wilms 2000 ; Christine et Wilms 2003 ; Christine 2006). L'échantillon-mère est obtenu par échantillonnage à plusieurs degrés autopondéré. Donc, les unités primaires sont tirées avec probabilités inégales qui sont proportionnelles à leurs tailles. Les variables d'équilibrage sont des variables sociodémographiques provenant du dernier recensement. Bardaji (2001) et Even (2002) ont également utilisé l'échantillonnage équilibré pour sélectionner un échantillon de bénéficiaires d'emplois subventionnés. Sept populations sont sondées, un échantillon équilibré de bénéficiaires est sélectionné dans chacune des populations en utilisant de deux à cinq variables d'équilibrage selon la population.

La société Electricité de France (EDF) a installé de nouveaux compteurs d'électricité de chaque ménage sur une base continue. La quantité d'informations recueillies est tellement grande qu'il est impossible d'archiver toutes les données. Dessertaine (2006, 2007) a utilisé l'échantillonnage équilibré pour sélectionner les séries chronologiques de données sur la consommation qui doivent être archivées afin de s'assurer qu'elles représentent aussi exactement que possible la consommation de l'ensemble de la population française. Biggert et Falorsi (2006) ont utilisé l'échantillonnage équilibré pour améliorer la qualité de l'indice des prix à la consommation en Italie. Gismondi (2007) a testé l'échantillonnage équilibré pour estimer le nombre de nuits que les touristes passent en Italie. D'Alò, Di Consiglio, Falorsi et Solari (2006) ainsi que Falorsi et Rigbi (2008) ont également proposé d'utiliser un plan de sondage équilibré pour estimer les totaux dans les petits domaines. Des simulations ont été exécutées par Mari, Barabà, Mitás et Passamonti (2007a, 2007b) en Argentine et par Chipperfield (2009) en Australie pour évaluer l'intérêt de l'échantillonnage équilibré pour l'échantillon-mère.

donc fort semblable à la variance estimée d'un estimateur par la régression généralisée (GREG). Néanmoins, la variance de l'estimateur GREG est généralement sous-estimée, parce qu'elle ne tient pas compte du caractère aléatoire des pondérations. En effet, si la variance habituelle de l'estimateur GREG est calculée pour le cas particulier de la poststratification, nous obtenons la variance d'un plan stratifié avec répartition proportionnelle. L'estimateur poststratifié est néanmoins plus grande que celle obtenue sous un plan stratifié avec répartition proportionnelle.

6.2 Approximation de la variance

Si le plan de sondage équilibré possède une grande entropie, Hájek (1981) ainsi que Deville et Tillé (2005, méthode 4) ont proposé l'approximation qui suit de la variance sous le plan donnée par :

$$\widehat{\text{var}}^p(\hat{Y}_\pi) \equiv \text{var}^{app}(\hat{Y}_\pi) = \sum_{k \in U} d_k \frac{(Y_k - \mathbf{x}_k' \mathbf{b})^2}{\pi_k^2}, \quad (5)$$

où l'indice inférieur p désigne le plan de sondage,

$$\mathbf{b} = \left(\sum_{k \in U} d_k \frac{\pi_k^2}{\mathbf{x}_k' \mathbf{x}_k'} \right)^{-1} \sum_{k \in U} d_k \frac{\pi_k^2}{\mathbf{x}_k' \mathbf{y}_k},$$

et les d_k sont la solution du système non linéaire

$$\pi_k(1 - \pi_k) = d_k - \left(\sum_{\ell \in U} d_\ell \frac{\pi_\ell^2}{\mathbf{x}_\ell' \mathbf{x}_\ell'} \right)^{-1} d_k \frac{\pi_k^2}{\mathbf{x}_k' \mathbf{x}_k'}, \quad k \in U. \quad (6)$$

L'entropie du plan de sondage dépend de la façon dont les vecteurs $\mathbf{u}(t)$ sont choisis durant la phase de vol. Afin d'accroître l'entropie, le vecteur $\mathbf{u}(t)$ peut être choisi aléatoirement ou bien la population peut être tirée aléatoirement avant de tirer l'échantillon. L'expression (5), qui ne contient que les probabilités d'inclusion de premier ordre, a été validée par Deville et Tillé (2005) sous une gamme d'échantillons équilibrés, indépendamment de la façon dont les valeurs de y étaient générées. Une approximation très proche de l'expression (5) a été obtenue par Fuller (2009) ainsi que par Legg et Yu (2010) pour un plan de sondage équilibré obtenu par une méthode réjective dans le cas d'un plan initial utilisant l'échantillonnage de Poisson. Ces approximations ne tiennent pas compte du problème d'arrondi.

6.3 Estimation de la variance

Deville et Tillé (2005) ont proposé une famille d'estimateurs de variance pour l'échantillonnage équilibré de la forme

où

$$\widehat{\text{var}}(\hat{Y}_\pi) = \sum_{k \in S} c_k \frac{(Y_k - \mathbf{x}_k' \mathbf{b})^2}{\pi_k^2}, \quad (7)$$

et les c_k sont les solutions du système non linéaire

$$1 - \pi_k = c_k - \left(\sum_{\ell \in S} c_\ell \frac{\pi_\ell^2}{\mathbf{x}_\ell' \mathbf{x}_\ell'} \right)^{-1} c_k \frac{\pi_k^2}{\mathbf{x}_k' \mathbf{x}_k'}, \quad (8)$$

qui peut être résolu au moyen d'un algorithme du point fixe. Dans Deville et Tillé (2005), des variantes plus simples de c_k ont également été proposées. Par exemple, on peut utiliser les valeurs de rechange,

$$\tilde{c}_k \approx \frac{n}{n - d} (1 - \pi_k),$$

qui sont très proches de c_k . L'estimateur $\widehat{\text{var}}(\hat{Y}_\pi)$ est approximativement sans biais sous le plan parce qu'il s'agit d'un estimateur par substitution de l'approximation donnée par l'expression (5) (pour plus de renseignements concernant les estimateurs obtenus par substitution, voir Deville 1999), qui est une approximation raisonnable de la variance sous le plan de sondage.

Il n'est pas facile d'utiliser la méthode du bootstrap pour estimer la variance dans le contexte de l'échantillonnage équilibré. Les échantillons équilibrés avec remise devaient être tirés de l'échantillon original. Une généralisation de la méthode du cube pour l'échantillonnage équilibré avec remise n'a pas encore été décrite. Une solution, proposée par Chauvet (2007), consiste à reconstruire une population artificielle d'après l'échantillon. Ensuite, des échantillons bootstrap sont tirés par une méthode d'échantillonnage équilibré. Une autre solution a été proposée par Fuller (2010) pour l'échantillonnage réjectif équilibré. Bredt et Chauvet (2010a) ont proposé une autre méthode dans laquelle une représentation de la méthode du cube par différence de martingale est utilisée pour approcher les probabilités d'inclusion de deuxième ordre, ce qui permet de construire un estimateur de variance presque sans biais.

7. Échantillonnage équilibré en pratique

7.1 Intérêt de l'échantillonnage équilibré

Dans les cadres assistés par modèle et basé sur un modèle, l'utilisation d'un plan de sondage équilibré avec l'estimateur de Horvitz-Thompson est souvent la stratégie optimale (voir Nedjalkova et Tillé 2009). En effet, quand l'échantillon est

effectuée aléatoirement de façon que $E[\pi(1)] = \pi(0)$. À la fin de la première étape de la phase de vol, nous avons donc atteint une face du cube, ce qui signifie qu'au moins une composante de $\pi(1)$ est égale à 0 ou à 1, c'est-à-dire que le problème est réduit d'un problème d'échantillonnage à partir d'une population de taille $N = 3$ à une population de taille $N = 2$. En N étapes au moins, la phase de vol est donc achevée.

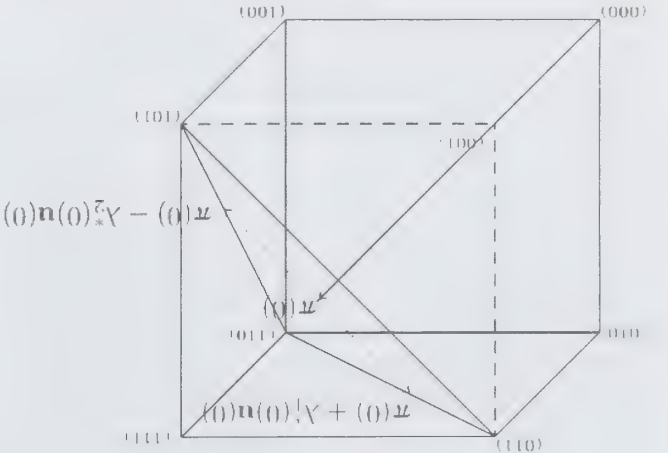


Figure 4 Phase de vol dans une population de taille $N = 3$ avec une contrainte de taille d'échantillon $n = 2$

De façon plus générale, la phase de vol est une marche aléatoire dans l'intersection du sous-espace d'équilibrage et du cube. Cette marche aléatoire s'arrête à un sommet de l'intersection du cube et du sous-espace. La phase de vol est définie par la classe suivante d'algorithmes. Commencer par initialiser à $\pi(0) = \pi$. Ensuite, au temps $t = 0, \dots, T$,

1. Générer un vecteur quelconque $\mathbf{u}(t) = [u_k(t)] \neq 0$ tel que
 - i) $\mathbf{u}(t)$ soit dans le noyau de la matrice $\mathbf{A} = (\mathbf{x}_1/\pi_1, \dots, \mathbf{x}_k/\pi_k, \dots, \mathbf{x}_N/\pi_N)$, c'est-à-dire $\mathbf{A}\mathbf{u}(t) = 0$,
 - ii) $u_k(t) = 0$ si $\pi_k(t)$ est un entier.
2. Calculer $\lambda_1^*(t)$ et $\lambda_2^*(t)$, les plus grandes valeurs telles que

$$0 \leq \pi(t) + \lambda_1^*(t)\mathbf{u}(t) \leq 1,$$

$$0 \leq \pi(t) - \lambda_2^*(t)\mathbf{u}(t) \leq 1.$$
3. Calculer

$$\pi(t+1) = \begin{cases} \pi(t) + \lambda_1^*(t)\mathbf{u}(t) & \text{avec la probabilité } q_1(t) \\ \pi(t) - \lambda_2^*(t)\mathbf{u}(t) & \text{avec la probabilité } q_2(t), \end{cases}$$

où $q_1(t) = \lambda_2^*(t)/\{\lambda_1^*(t) + \lambda_2^*(t)\}$ et $q_2(t) = 1 - q_1(t)$.

La phase de vol s'arrête quand il n'est plus possible de trouver un vecteur $\mathbf{u}(t) \neq 0$.

5.3 Phase d'atterrissage

Si, à la fin de la phase de vol, les équations d'équilibrage ne sont pas exactement satisfaites, la phase d'atterrissage est nécessaire. Soit $\pi^* = [\pi_k^*]$ le vecteur obtenu à la dernière étape de la phase de vol. Il est possible de prouver (voir Deville et Tillé 2004) que

$$\text{card}(U^*) \leq p,$$

où

$$U^* = \{k \in U \mid 0 < \pi_k^* < 1\}$$

et p est le nombre de variables d'équilibrage. Le but de la phase d'atterrissage est de trouver un échantillon \mathbf{s} tel que $E(\mathbf{s}|\pi^*) = \pi^*$, qui est presque équilibré. Il existe deux moyens de tirer un tel échantillon :

1. La phase de vol par programmation linéaire consiste à considérer tous les échantillons possibles de U^* . Un coût est attribué à chaque échantillon. Ce coût est, par exemple, la distance entre l'échantillon et le sous-espace des contraintes. Ensuite, on recherche un plan de sondage de U^* qui minimise le coût prévu et qui satisfait les probabilités d'inclusion π^* . Ce problème peut être résolu parce que le nombre d'échantillons à considérer est raisonnable étant donné la petite taille de U^* .

2. La phase de vol par suppression de variables peut être utilisée quand le nombre de variables d'équilibrage est trop grand pour que le problème de programmation linéaire puisse être résolu par l'algorithme du simplexe ($p > 20$). Si l'on applique cette méthode, une variable auxiliaire est abandonnée à la fin de la phase de vol. Ensuite, on peut retourner à la phase de vol jusqu'à ce qu'il ne soit plus possible de contraintes sont alors relâchées successivement selon un ordre de préférence.

6. Variance et estimation de la variance

6.1 Une technique de résidu

La variance de l'estimateur de Horvitz-Thompson peut être estimée en appliquant une technique élaborée dans Deville et Tillé (2005). Cette technique est comparable à celle utilisée pour estimer la variance de l'estimateur par calage et a été validée par un ensemble de simulations. La variance estimée de l'estimateur de Horvitz-Thompson est

$\pi = (0,5, 0,5, 0,5, 0,5)$ et $n = 2$, nous sommes capables d'obtenir la série suivante de vecteurs :

$$\pi = \begin{pmatrix} 0,5 \\ 0,5 \\ 0,5 \\ 0,5 \\ 0,5 \\ 0,5 \end{pmatrix} \rightarrow \begin{pmatrix} 0 \\ 0,6666 \\ 0,6666 \\ 0,6666 \\ 0,6666 \\ 0 \end{pmatrix} \rightarrow \begin{pmatrix} 0 \\ 0,5 \\ 0,5 \\ 0,5 \\ 0,5 \\ 0 \end{pmatrix} \rightarrow \begin{pmatrix} 0 \\ 1 \\ 1 \\ 1 \\ 1 \\ 0 \end{pmatrix} = s.$$

L'algorithme s'arrête quand toutes les composantes du vecteur sont égales à 0 ou 1.

Exemple 3. Si la contrainte est la taille fixe d'échantillon, un problème d'arrondi n'est pas un entier. En cas de problème d'arrondi, certaines composantes ne peuvent pas être fixées à zéro. Par exemple, avec $\pi = (0,5, 0,5, 0,5, 0,5, 0,5)$ et

$$\sum_{k \in U} \pi_k = 2,5,$$

nous pouvons observer la série suivante de vecteurs :

$$\pi = \begin{pmatrix} 0,5 \\ 0,5 \\ 0,5 \\ 0,5 \\ 0,5 \\ 0,5 \end{pmatrix} \rightarrow \begin{pmatrix} 0,625 \\ 0 \\ 0,625 \\ 0,625 \\ 0,625 \\ 0,625 \end{pmatrix} \rightarrow \begin{pmatrix} 0,5 \\ 0 \\ 0,5 \\ 0,5 \\ 0,5 \\ 0,5 \end{pmatrix} \rightarrow \begin{pmatrix} 0,5 \\ 0 \\ 0,25 \\ 0,25 \\ 0,25 \\ 0,25 \end{pmatrix} \rightarrow \begin{pmatrix} 0,5 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix} \rightarrow \begin{pmatrix} 1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix} = \pi^*.$$

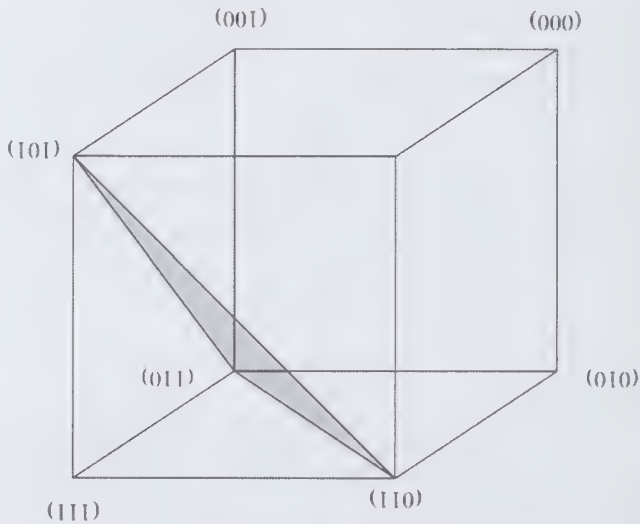
Dans ce cas, la phase de vol ne peut pas se terminer par un vecteur de 0 ou de 1 dont la somme est égale à 2,5. Dans ces conditions, la phase de vol se termine par un vecteur contenant une composante non entière.

5.2 La phase de vol

La première étape de la phase de vol est présentée à la figure 4 pour un cas très particulier : la taille de population $N = 3$. La seule contrainte d'équilibrage est la taille fixe d'échantillon $n = 2$. À la première étape, un vecteur $u(0)$ doit être choisi. Il peut l'être librement, mais doit être tel que $\pi + u(0)$ demeure dans le sous-espace des contraintes. En fait, la méthode du cube correspond à une famille de méthodes qui dépendent de la façon dont le vecteur $u(0)$ est choisi. Il peut l'être aléatoirement ou non.

Si, en partant de π , nous suivons la direction donnée par le vecteur $u(0)$, nous aboutissons nécessairement à une face du cube. Considérons ce point désigné sur la figure 4 par $\pi(0) + \lambda_1^*(0)u(0)$. Maintenant si, en partant de π , nous suivons la direction opposée, c'est-à-dire la direction donnée par le vecteur $-u(0)$, nous aboutissons sur la figure 4 par $\pi(0) - \lambda_2^*(0)u(0)$. À la première étape, le vecteur $\pi(0) = \pi(0) + \lambda_1^*(0)u(0)$ ou à $\pi(0) - \lambda_2^*(0)u(0)$. Le choix est

Figure 2 Échantillons possibles dans une population de taille $N = 3$ avec une contrainte de taille d'échantillon fixe $n = 2$



et du sous-espace de contrainte. Cette marche aléatoire s'arrête à un sommet de l'intersection du cube et du sous-espace des contraintes. À la fin de la phase de vol, si aucun échantillon n'a été obtenu, la phase d'atterrissage se déclenche en vue de sélectionner un échantillon aussi proche que possible du sous-espace des contraintes.

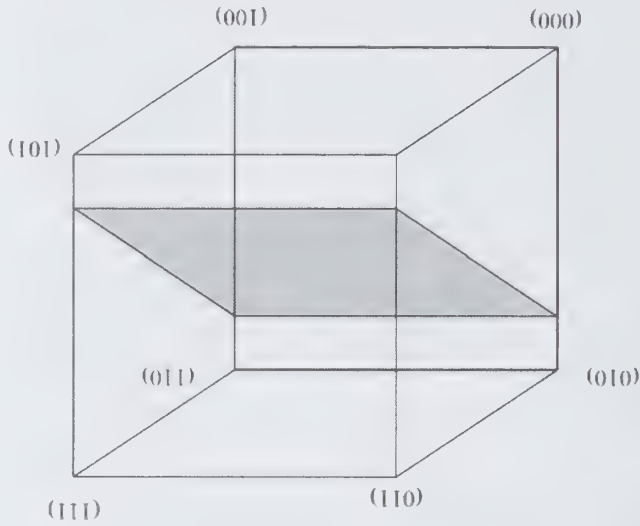


Figure 3 Échantillons possibles dans une population de taille $N = 3$ avec une contrainte et un problème d'arrondi

Exemple 2. Si la contrainte est la taille fixe d'échantillon, la phase de vol transforme aléatoirement un vecteur de probabilités d'inclusion en un vecteur de 0 et de 1. À chaque étape de l'algorithme, le vecteur de probabilités d'inclusion est transformé aléatoirement, mais la somme des probabilités d'inclusion doit demeurer égale à n . Par exemple, avec

Cette méthode est générale en ce sens que les probabilités d'inclusion sont exactement satisfaites, que ces probabilités peuvent être égales ou inégales, et que l'échantillon est aussi équilibré que possible.

Fuller (2009) a étudié une procédure réjective en fixant un intervalle de tolérance en dehors duquel l'échantillon est rejeté et a proposé un estimateur de variance. Même si les probabilités d'inclusion sont modifiées par une procédure réjective, Fuller (2009) montre que des estimations efficaces sont obtenues en utilisant les probabilités d'inclusion du plan original. En utilisant un ensemble de simulations, Legg et Yu (2010) ont comparé cette procédure réjective à la méthode du cube et montre que les deux méthodes donnent des résultats équivalents. Enfin, Dudoignon et Vanheuverzwyn (2006) ont proposé une méthode rapide d'échantillonnage équilibré pour les totaux marginaux, tandis que Périé (2008) a proposé une méthode basée sur des nombres aléatoires permanents qui fournissent un échantillon équilibré. Dans le cas de la méthode de Périé (2008), les probabilités d'inclusion ne sont qu'approximativement satisfaites.

5. La méthode du cube

5.1 Idées principales

La méthode du cube (voir Deville et Tillé 2004 ; Tillé 2001, 2006a, 2006b ; Ardilly 2006) représente une classe d'algorithme d'échantillonnage qui réalise le tirage d'un échantillon équilibré et satisfait exactement un ensemble de probabilités d'inclusion données. La méthode du cube est une extension de la méthode de scission élaborée par Deville et Tillé (1998). Elle est basée sur une transformation aléatoire du vecteur des probabilités d'inclusion jusqu'à l'obtention d'un échantillon tel que :

- (i) les probabilités d'inclusion soient exactement satisfaites ;
- (ii) les équations d'équilibrage soient satisfaites autant qu'il est possible.

La méthode doit son nom à la représentation géométrique d'un plan de sondage. En effet, un échantillon peut être représenté par un vecteur d'indicateurs d'échantillons :

$$s = (I[1 \in s] \dots I[k \in s] \dots I[N \in s]),$$

où $I[k \in s]$ prend la valeur 1 si $k \in s$ et 0 autrement. Un échantillon peut donc être vu comme l'un des sommets d'un hypercube de dimension N comme l'illustre la figure 1.

Figure 1 Échantillons possibles dans une population de taille $N = 3$

Définissons aussi

$$E(s) = \sum_{s \in S} p(s)s = \pi,$$

où $\pi = [\pi_k]$ est le vecteur des probabilités d'inclusion. Les équations d'équilibrage

$$\sum_{k \in S} \frac{\mathbf{x}_k}{\pi_k} = \sum_{k \in U} \mathbf{x}_k \pi_k,$$

peuvent aussi s'écrire

$$\sum_{k \in U} \mathbf{x}_k s_k = \sum_{k \in U} \mathbf{x}_k \pi_k, \quad (4)$$

où $s_k \in \{0, 1\}$ et $\mathbf{x}_k = \mathbf{x}_k / \pi_k$, $k \in U$. L'expression (4) est un système d'équations contenant des valeurs inconnues s_k qui définissent un sous-espace affine dans \mathbb{R}^N de dimension $N - p$ désigné par \tilde{Q} , où

$$\tilde{Q} = \left\{ \mathbf{u} \in \mathbb{R}^N \mid \sum_{k \in U} \mathbf{x}_k u_k = \sum_{k \in U} \mathbf{x}_k \right\}.$$

Le problème du tirage d'un échantillon équilibré peut donc être reformulé. Un plan de sondage équilibré consiste à choisir un sommet de l'hypercube de dimension N (un échantillon) qui demeure dans le sous-espace linéaire \tilde{Q} . Les figures 2 et 3 montrent, respectivement, deux exemples : le premier correspond à une contrainte de taille d'échantillon fixe et le deuxième, à une contrainte donnant lieu à un problème d'arrondi.

La méthode du cube (Deville et Tillé 2004) est divisée en deux phases : la phase de vol et la phase d'atterrissage. La phase de vol est une marche aléatoire qui part du vecteur des probabilités d'inclusion et demeure à l'intersection du cube

d'unités dans l'échantillon jusqu'à ce qu'un échantillon suffisamment équilibré soit obtenu.

Dans le cadre basé sur un modèle, Royall (1976a, 1976b) a préconisé d'utiliser l'échantillonnage équilibré afin d'atteindre la stratégie optimale et de se protéger contre l'erreur de spécification du modèle (voir aussi Royall et Pfeffermann 1982; Kott 1986; Cumberland et Royall 1988; Royall 1988; Tirat 2006; Nediyalkova et Tillé 2009). Bien que plusieurs méthodes de sélection d'un échantillonnage équilibré soient présentées dans le livre de Valliant, Dorfman et Royall (2000), ces méthodes ne spécifient pas nécessairement les probabilités d'inclusion de l'échantillon. Dans le cadre basé sur un modèle, il est important que l'échantillon soit équilibré. Cependant, cet échantillon ne doit pas toujours être sélectionné aléatoirement.

Hájek (1981) a également recommandé d'utiliser l'échantillonnage équilibré. Pour Hájek, un échantillonnage équilibré est un cas particulier d'une stratégie représentative, une stratégie consistant en un couple formé d'un plan de sondage et d'un estimateur. Une stratégie représentative est une stratégie qui permet d'estimer les totaux des variables auxiliaires sans erreur. En ce sens, le plan de sondage équilibré avec l'estimateur de Horvitz-Thompson est une stratégie représentative. Hájek (1981) propose une procédure réjective qui consiste à tirer une série d'échantillons jusqu'à ce que l'on obtienne un échantillon équilibré. Les procédures réjectives ont deux inconvénients : si plusieurs variables d'équilibrage sont utilisées, la procédure peut être très lente ; deuxièmement, les probabilités d'inclusion des plan réjects ne sont pas les mêmes que celle du plan original. Les probabilités d'inclusion des unités statistiques qui sont proches des moyennes de population sont augmentées au détriment de celles des unités qui sont éloignées du centre (voir par exemple les simulations de Legg et Yu 2010).

Une autre méthode de sélection consiste à énumérer tous les échantillons possibles, puis à construire un plan de sondage uniquement pour tirer les échantillons qui sont adéquatement équilibrés. Un plan de ce genre peut être construit en recourant à la programmation linéaire. Cette technique a été utilisée par Ardilly (1991) pour sélectionner les unités primaires de l'échantillon-maître français. Néanmoins, cette méthode ne s'applique qu'à de petites tailles de population en raison de l'explosion combinatoire du nombre d'échantillons quand la taille de la population est grande.

Déville, Grosbras et Roth (1988) et Deville (1992) ont proposé des méthodes multivariées d'échantillonnage équilibré avec probabilités d'inclusion égales. Hedayat et Majumdar (1995) ont proposé l'adaptation d'une technique basée sur un plan expérimental qui permettrait de créer un plan de sondage équilibré. De nouveau, cette technique est limitée aux probabilités d'inclusion égales. Enfin, la méthode du cube a été proposée par Deville et Tillé (2004).

3.3 Équilibrage sur une constante

sélectionné dans chacune d'elles. Dans le contexte du calage, les statisticiens effectuent généralement le calage sur les totaux marginaux et non sur toutes les cellules contenues dans un tableau de contingence. Puisqu'un échantillonnage équilibré peut être considéré comme une sorte de calage qui est intégré directement dans le plan de sondage, on souhaiterait également effectuer l'équilibrage en utilisant uniquement les totaux marginaux. Néanmoins, la théorie habituelle de la stratification ne permet pas le chevauchement des strates, puisque la stratification doit être une partition de la population. Maintenant, la méthode du cube permet d'équilibrer directement sur les totaux des strates chevauchantes en utilisant simplement les indicateurs des strates comme variables d'équilibrage.

Un autre cas particulier intéressant de l'échantillonnage équilibré est celui où une constante est utilisée comme variable d'équilibrage. Si $\mathbf{x}_k = 1$ pour tout $k \in U$, les équations d'équilibrage deviennent

$$\sum_{k \in U} \frac{1}{\pi_k} = \sum_{k \in U} 1 = N.$$

En fait,

$$\sum_{k \in S} \pi_k$$

est l'estimateur de Horvitz-Thompson de N . Cela signifie que, si la variable d'équilibrage utilisée est une constante, la taille de population estimée concorde avec la taille connue N , ce qui est loin d'être un fait acquis quand les unités statistiques sont sélectionnées avec probabilités d'inclusion inégales.

4. Historique du concept d'équilibrage et méthodes existantes

L'idée de l'échantillonnage équilibré est très ancienne et reliée au vague concept de représentativité qui était déjà employé par Kiaer (1896, 1899, 1903, 1905). Le premier article consacré au tirage d'un échantillon équilibré est dû à Gini (1928) et à Gini et Galvani (1929) qui ont tiré un échantillon de 29 parmi 214 districts italiens afin d'égaliser plusieurs totaux de population. Neyman (1952) et Yates (1960) ont tous deux condamné l'article de Gini et Galvani, essentiellement parce que la sélection de l'échantillon n'était pas aléatoire (voir Langel et Tillé 2010). Les premières méthodes de tirage d'un échantillon équilibré aléatoire ont été proposées par Yates (1946) et par Thionet (1953), mais elles étaient réjectives en ce sens qu'elles comportaient le tirage aléatoire d'échantillons ou le remplacement aléatoire

3. Cas particuliers d'échantillonnage équilibré

3.1 Échantillonnage avec probabilités inégales et stratification

Certains plans d'échantillonnage bien connus sont des cas particuliers d'échantillonnage équilibré :

1. L'échantillonnage avec une taille d'échantillon fixe est un cas particulier de l'échantillonnage équilibré. Dans ce cas, la seule variable d'équilibrage est π_k . Les équations d'équilibrage données en (2) deviennent

$$\sum_{k \in S} \frac{\pi_k}{\pi_k} = \sum_{k \in S} 1 = \sum_{k \in U} \pi_k,$$

ce qui signifie que la taille d'échantillon doit être fixe.

2. La stratification est un cas particulier d'échantillonnage équilibré. Supposons que la population est partitionnée en H strates $U_h, h = 1, \dots, H$, de tailles $N_h, h = 1, \dots, H$, et qu'un échantillon est tiré dans chaque strate par échantillonnage aléatoire simple sans remise avec taille d'échantillon fixe $n_h, h = 1, \dots, H$. Dans ce cas, les variables d'équilibrage sont les variables indicatrices des strates

$$\delta_{kh} = \begin{cases} 1 & \text{si } k \in U_h \\ 0 & \text{autrement.} \end{cases}$$

Sous un plan de sondage stratifié, les estimateurs de Horvitz-Thompson des tailles des strates sont exactement égaux aux tailles des strates, ce qui est une propriété de l'équilibrage sur les variables indicatrices des strates. En effet, puisque les probabilités d'inclusion dans la strate h sont $\pi_k = n_h / N_h, k \in U_h$, les équations d'équilibrage deviennent

$$\sum_{k \in S} \frac{N_h \delta_{kh}}{n_h} = \sum_{k \in U} \delta_{kh} = N_h, h = 1, \dots, H,$$

et sont exactement satisfaites.

Ces deux plans sont bien connus et sont appliqués fréquemment en statistique officielle afin de réduire la variance. Le concept plus général d'équilibrage donne plus de liberté en vue de choisir les meilleures variables d'équilibrage qui augmenteront l'exactitude des estimateurs.

3.2 Strates chevauchantes

La construction d'un plan de sondage stratifié est souvent un exercice difficile. Les statisticiens essaient fréquemment d'effectuer la stratification en utilisant plusieurs variables qualitatives. Cependant, dans la plupart des cas, le croisement de toutes les strates de toutes les variables rend les cellules trop petites pour qu'un échantillon puisse être

2.2 Plan de sondage équilibré

Autrement dit, dans un échantillon équilibré, le total des variables x est estimé sans erreur. Plusieurs auteurs, dont Cumberland et Royall (1981) et Kott (1986), donneraient à un échantillon qui satisfait l'équation (2) le nom d'« échantillon équilibré sur la moyenne » défini par l'équation (1). Néanmoins, dans le présent article, nous considérerons que (1) est simplement un cas particulier de (2) qui se produit quand $\pi_k = n/N$ ou quand l'échantillon n'est pas tiré aléatoirement. Dans les deux cas, nous parlons d'un échantillon équilibré.

Soit $p(s)$ le plan de sondage, c'est-à-dire la probabilité

que l'échantillon s soit sélectionné, telle que $p(s) = \Pr(S = s)$, où S est l'échantillon aléatoire et $n(S)$, la taille de l'échantillon S . Selon la définition de Deville et Tiïlle (2004), un plan de sondage $p(\cdot)$ est dit *équilibré* sur les variables auxiliaires x_1, \dots, x_p si l'estimateur de Horvitz-Thompson satisfait l'équation (2). Dans un plan de sondage équilibré, les probabilités d'inclusion sont fixées avant le tirage de l'échantillon. Un échantillonnage équilibré peut être considéré comme une sorte de calage qui est directement intégré dans le plan de sondage. Le principal problème est que les équations d'équilibrage (2) peuvent rarement être satisfaites exactement. Nous donnons à cette difficulté le nom de « problème d'arrondi ».

Exemple 1. Si $N = 4, n = 2, \pi_k = 1/2$, pour tout $k \in U$ et $x_1 = 0, x_2 = 1, x_3 = 2, x_4 = 4$, les équations d'équilibrage

$$\frac{1}{n} \sum_{k \in S} x_k = \frac{1}{N} \sum_{k \in U} x_k,$$

qui équivaut à

$$\sum_{k \in S} x_k = \frac{N}{n} \sum_{k \in U} x_k. \quad (3)$$

Puisque

$$\frac{N}{n} \sum_{k \in U} x_k = \frac{4}{2} (0 + 1 + 2 + 4) = 3.5,$$

et que le premier membre de (3) est toujours un nombre entier, un échantillon exactement équilibré n'existe pas.

Effectivement, la sélection de l'échantillon est un problème en nombres entiers. La méthode du cube a donc pour objectif de tirer un échantillon qui satisfait exactement les probabilités d'inclusion π_k tout en restant aussi équilibré que possible.

Dix années d'échantillonnage équilibré par la méthode du cube : une évaluation

Yves Tillé¹

Résumé

Le présent article propose un examen et une évaluation de l'échantillonnage équilibré par la méthode du cube. Il débute par une définition de la notion d'échantillon équilibré et d'échantillonnage équilibré, suivie par un court historique du concept d'équilibrage. Après un exposé succinct de la théorie de la méthode du cube, l'accent est mis sur les aspects pratiques de l'échantillonnage équilibré, c'est-à-dire l'intérêt de la méthode comparative à d'autres méthodes d'échantillonnage et au calage, le domaine d'application, la précision de l'équilibrage, le choix des variables auxiliaires et les moyens de mettre la méthode en œuvre.

Mots clés : Échantillonnage ; équilibrage ; estimateur de Horvitz-Thompson.

1. Introduction

Bien que le concept d'échantillonnage équilibré existe depuis les tous débuts de la statistique d'enquête, son application a été difficile, parce que presque toutes les méthodes proposées étaient énumératives ou réjectives et que le temps de calcul était considérable. L'algorithme de la méthode du cube a été proposé en 1998 par Deville et Tillé, et une première implémentation a été écrite par trois étudiants de l'École Nationale de la Statistique et de l'Analyse de l'Information de Rennes, en France (voir Bousabaa, Lieber et Sirouil 1999). Finalement, la méthode a été publiée dans Tillé (2001) et dans Deville et Tillé (2004). Depuis, plusieurs implémentations de la méthode du cube ont été proposées et plusieurs gestionnaires d'enquête l'ont utilisée pour sélectionner des échantillons, les applications les plus importantes étant le nouveau recensement français et l'échantillon-matrice français.

Notre objectif est d'évaluer le développement et l'utilisation de l'échantillonnage équilibré aux cours des dix dernières années afin de mieux déterminer quand et comment la méthode du cube peut être appliquée pour sélectionner les échantillons de ménages ou d'établissements. À la section 2, nous discutons du concept d'échantillon équilibré et d'échantillonnage équilibré. À la section 3, nous présentons une liste de cas particuliers. À la section 4, nous faisons succinctement l'historique de ce concept pour le cadre basé sur un modèle et celui basé sur le plan. Ensuite, à la section 5, nous donnons un bref aperçu de la méthode du cube, qui représente une classe d'algorithmes permettant de sélectionner aléatoirement des échantillons équilibrés en fixant les probabilités d'inclusion (voir Deville et Tillé 2004 ; Tillé 2001, 2006b). Nous tentons de présenter les grands principes de cet algorithme sans nous attarder à la description détaillée des aspects purement techniques de la méthode. La

section 6 est consacrée aux principes d'estimation de la variance sous échantillonnage équilibré. Enfin, à la section 7, nous discutons de l'intérêt pratique de l'échantillonnage équilibré et comparons ce dernier à d'autres méthodes d'échantillonnage et au calage. Nous donnons également une liste d'applications récentes. Cette section traite aussi de l'exactitude de l'équilibrage, du choix des variables auxiliaires et des moyens de mettre en œuvre l'échantillonnage équilibré. L'article se termine par une bibliographie exhaustive sur l'échantillonnage équilibré et ses applications.

2. Échantillonnage équilibré

2.1 Définition d'un échantillon équilibré

Considérons un échantillon s de taille n qui est un sous-ensemble d'une population finie U de taille N . Un échantillon est dit équilibré si, pour un vecteur de variables auxiliaires $\mathbf{x}_k = (x_{k1}, \dots, x_{kj}, \dots, x_{kp})'$,

$$(1) \quad \frac{1}{n} \sum_{k \in s} \mathbf{x}_k = \frac{1}{N} \sum_{k \in U} \mathbf{x}_k$$

ce qui signifie que les moyennes d'échantillon des variables x concordent avec leurs moyennes de population.

Brewer (1999) fait une distinction entre la sélection équilibrée d'échantillons et la sélection aléatoire d'échantillons. Cependant, un échantillon équilibré peut être sélectionné aléatoirement. Si un échantillon aléatoire S est sélectionné aléatoirement, chaque unité de la population a une probabilité π_k d'inclusion dans l'échantillon. Dans ce cas, un échantillon aléatoire doit satisfaire les équations d'équilibrage suivantes :

$$(2) \quad \sum_{k \in S} \pi_k \mathbf{x}_k = \sum_{k \in U} \mathbf{x}_k$$

- Lepkowski, J.M., et Groves, R.M. (1986). A mean squared error model for multiple frame, mixed mode survey design. *Journal of the American Statistical Association*, 81, 930-937.
- Lesser, V.M., et Kaltsbeek, W.D. (1999). Nonsampling errors in environmental surveys. *Journal of Agricultural, Biological, and Environmental Statistics*, 4, 473-488.
- Lohr, S.L. (2007). Recent developments in multiple frame surveys. *Proceedings of the Survey Research Methods Section, American Statistical Association*, 3257-3264.
- Lohr, S.L. (2009). Multiple frame surveys. Dans *Handbook of Statistics, Sample Surveys: Design, Methods and Applications*, (Éds., D. Pfeffermann et C.R. Rao). Amsterdam : North Holland, Vol. 29A, 71-88.
- Lohr, S.L., et Rao, J.N.K. (2000). Inference in dual frame surveys. *Journal of the American Statistical Association*, 95, 271-280.
- Lohr, S.L., et Rao, J.N.K. (2006). Estimation in multiple-frame surveys. *Journal of the American Statistical Association*, 101, 1019-1030.
- Lu, W., Brick, J.M. et Sitter, R. (2006). Algorithms for constructing combined strata variance estimators. *Journal of the American Statistical Association*, 101, 1680-1692.
- Lu, Y., et Lohr, S.L. (2010). L'estimation des flux bruts dans les enquêtes à base de sondage double. *Techniques d'enquête*, 36, 13-24.
- Mecatti, F. (2007). Un estimateur à base de sondage unique fondé sur la multiplicité pour les sondages à bases multiples *Techniques d'enquête*, 33, 171-178.
- Rao, J.N.K. (2003). *Small Area Estimation*. New York : John Wiley & Sons, Inc.
- Rao, J.N.K., et Wu, C. (2010). Pseudo-empirical likelihood inference for dual frame surveys. *Journal of the American Statistical Association*, 105, 1494-1503.
- Skinner, C.J. (1991). On the efficiency of raking ratio estimation for multiple frame surveys. *Journal of the American Statistical Association*, 86, 779-784.
- Skinner, C.J., Holmes, D.J. et Holt, D. (1994). Multiple frame sampling for multivariate stratification. *Revue Internationale de Statistique*, 62, 333-347.
- Skinner, C.J., et Rao, J.N.K. (1996). Estimation in dual frame surveys with complex designs. *Journal of the American Statistical Association*, 91, 349-356.
- Thompson, S.K. (2002). *Sampling Techniques*, 2^e Ed. New York : John Wiley & Sons, Inc.
- Vannieuwenhuyze, J., Loosveldt, G. et Moltenberghs, G. (2011). A method for evaluating mode effects in mixed-mode surveys. *Public Opinion Quarterly*, 74, 1027-1045.

bases de sondage multiples peut, comme les autres enquêtes, présenter une non-réponse, des effets de mode et des erreurs de mesure. En outre, à moins que toutes les bases de sondage correspondent à la population complète, les estimateurs pour enquête à bases de sondage multiples peuvent être sensibles à l'erreur de classification dans les domaines. Une correction pour l'erreur de classification a été donnée à la section 6, mais une étude plus approfondie de ces défis est nécessaire. Les effets de l'erreur de classification dans les domaines, de la non-réponse et du biais de mode peuvent être confondus. Une expérience peut aider à débrouiller ces effets. Nous étudions à l'heure actuelle la relation entre ces trois types d'erreurs non liées à l'échantillonnage. Chaque forme d'erreur non due à l'échantillonnage a une incidence sur l'exactitude des estimateurs pour bases de sondage multiples, et les erreurs non dues à l'échantillonnage attendues doivent être intégrées dans un plan de sondage optimal.

Remerciements

Cette étude a été financée en partie par la National Science Foundation aux termes des subventions SES-0604373 et DRL-0909630. Une partie de la matière du présent article a été présentée à l'assemblée annuelle de 2010 de la Société de statistique de l'Italie et publiée dans les actes de cette conférence. L'auteur remercie les examinateurs de leurs commentaires utiles.

Bibliographie

Bankier, M.D. (1986). Estimators based on several stratified samples with applications to multiple frame surveys. *Journal of the American Statistical Association*, 81, 1074-1079.

Biemer, P.P. (1984). Methodology for optimal dual frame sample design. Bureau of the Census SRD Research Report CENSUS/SRD/RR-84/07.

Brick, J.M. (2010). Dual frame landline and cell phone surveys. Article présenté à l'Annual meeting of the Statistical Society of Canada, Québec.

Brick, J.M., Cervantes, I.F., Lee, S. et Norman, G. (2011). Erreurs non dues à l'échantillonnage dans les enquêtes téléphoniques à base de sondage double. *Techniques d'enquête*, 37, 1, 1-16.

Brick, J.M., Dipko, S., Presser, S., Tucker, C. et Yuan, Y. (2006). Nonresponse bias in a dual frame survey of cell and landline numbers. *Public Opinion Quarterly*, 70, 780-793.

Chambers, R., Chipperfield, J., Davis, W. et Kovacevic, M. (2008). Inference based on estimating equations and probability-linked data. University of Wollongong Centre for Statistical & Survey Methodology Working Paper 18-09.

Citro, C.F., et Kalton, G., Eds. (2007). *Using the American Community Survey: Benefits and Challenges*. Washington, D.C.: National Academies Press.

Clark, J., Winglee, M. et Lin, B. (2007). Handling imperfect overlap determination in a dual-frame survey. *Proceedings of the Survey Research Methods Section, American Statistical Association*, 3233-3238.

Cochran, W.G. (1977). *Sampling Techniques*, 3^e Ed. New York : John Wiley & Sons, Inc.

de Leeuw, E. (2008). Choosing the method of data collection. Dans *International Handbook of Survey Methodology*, (Eds., E. de Leeuw, J. Hox et D. Dillman). New York : Lawrence Erlbaum, 113-135.

de Leeuw, E., Hox, J. et Dillman, D. (2008). Mixed-mode surveys: When and why. Dans *International Handbook of Survey Methodology*, (Eds., E. de Leeuw, J. Hox et D. Dillman). New York : Lawrence Erlbaum, 299-316.

Dernath, A., Rao, J.N.K., Hidiroglou, M.A. et Tamby, J.-L. (2007). On the allocation and estimation for dual frame survey data. *Proceedings of the Survey Research Methods Section, American Statistical Association*, 2938-2945.

Deville, J.-C., et Samdal, C.-E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, 87, 376-382.

Fuller, W.A., et Burmeister, L.F. (1972). Estimators for samples selected from two overlapping frames. *Proceedings of the Social Statistics Section, American Statistical Association*, 245-249.

Gonzalez-Villalobos, A., et Wallace, M.A. (1996). *Multiple Frame Agriculture Surveys*. Rome : Food and Agriculture Organization of the United Nations. Vols. 1 et 2.

Haines, D.E., et Pollock, K.H. (1998). Combinaison de bases multiples pour estimer la taille et les chiffres de la population. *Techniques d'enquête*, 24, 81-91.

Hansen, M.H., Hurwitz, W.N. et Madow, W.G. (1953). *Sample Survey Methods and Theory*. New York : John Wiley & Sons, Inc. Volume 1.

Hartley, H.O. (1962). Multiple frame surveys. *Proceedings of the Social Statistics Section, American Statistical Association*, 203-206.

Hartley, H.O. (1974). Multiple frame methodology and selected applications. *Sankhyā, Séries C*, 36, 99-118.

Horvitz, D.G., et Thompson, D.J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47, 663-685.

Iachan, R., et Dennis, M.L. (1993). A multiple frame approach to sampling the homeless and transient population. *Journal of Official Statistics*, 9, 747-764.

Kalton, G., et Anderson, D.W. (1986). Sampling rare populations. *Journal of the Royal Statistical Society, Séries A*, 149, 65-82.

Kennedy, C. (2007). Evaluating the effects of screening for telephone service in dual frame RDD surveys. *Public Opinion Quarterly*, 71, 750-771.

Kott, P.S., Amrhein, J.F. et Hicks, S.D. (1998). Échantillonnage et estimation à partir de bases de sondage listes multiples. *Techniques d'enquête*, 24, 3-10.

indépendants provenant de la strate ab . Le problème de plan de sondage peut être approché comme un plan de sondage stratifié.

En général, le plan optimal est une fonction des variances d'échantillonnage et des erreurs non dues à l'échantillonnage dans chaque base de sondage, ainsi que de l'estimateur choisi. Biemer (1984) ainsi que Lepkowski et Groves (1986) ont discuté de plans de sondage pour la situation de la figure 1 quand un échantillon à plusieurs degrés est tiré de chaque base de sondage en utilisant l'estimateur de Hartley $\bar{Y}(\theta^H)$. Lepkowski et Groves (1986) ont tenu compte de la variabilité due à l'intervieweur et du biais dû au mode, ainsi que de l'erreur d'échantillonnage pour évaluer la précision des divers plans ; de plus grandes tailles d'échantillons sont affectées aux bases de sondage dont le biais de mode est plus faible. Brick (2010) a obtenu les répartitions optimales en présence de non-réponse et constaté que tenir compte de la non-réponse peut affecter les ressources aux deux bases de sondage peut accroître considérablement l'efficacité dans les enquêtes à base de sondage double avec sélection ainsi qu'avec chevauchement.

L'un des avantages d'un plan à bases de sondage multiples est sa souplesse ; il convient bien pour une approche modulaire du plan de sondage. Dans certaines situations, il pourrait être pratique de tirer un échantillon initial de la population générale (base de sondage A dans la figure 4). La conception des échantillons provenant des bases de sondage B et C, qui correspondent à des sous-populations d'intérêt, peut alors être déterminée en se servant de l'information figurant dans l'échantillon de la base de sondage A. Par exemple, si l'échantillon provenant de la base de sondage A produit un trop petit nombre d'ingénieurs, la taille de l'échantillon tiré d'une liste de membres d'une association d'ingénieurs peut être augmentée en conséquence.

Rao (2003) a proposé d'utiliser des enquêtes à bases de sondage multiples pour améliorer l'exactitude des estimations sur petits domaines dans les sous-groupes d'intérêt. Dans cette application, des enquêtes supplémentaires peuvent être réalisées à l'aide de bases de sondage à forte concentration des sous-groupes d'intérêt. À mesure que les besoins de recherche évoluent, les ressources peuvent être réaffectées entre les enquêtes supplémentaires sans changer le plan de l'enquête principale. Une enquête sur la victimisation s'appuyant sur une base areolaire nationale peut être complétée par des enquêtes locales sur la victimisation ; à mesure que les profils de victimisation changent, les enquêtes locales peuvent être réalisées auprès d'échantillons de diverses tailles ou être transférées vers d'autres régions géographiques.

8. Conclusion

Dans le présent article, nous avons résumé un problème que pose l'utilisation des méthodes à bases de sondage multiples pour les enquêtes-ménages américaines. Les plans à bases de sondage multiples offrent d'énormes possibilités d'améliorer l'efficacité de la collecte des données dans les enquêtes-ménages. Ils peuvent accroître la couverture en combinant des bases de sondage incomplètes, améliorer l'exactitude des estimations pour des sous-groupes ou populations rares, et accroître la souplesse et la capacité d'adaptation de la collecte des données fédérales. Les enquêtes à bases de sondage multiples peuvent faciliter l'échantillonnage de populations difficiles à joindre, tels que les nouveaux immigrants ou les ménages avec des nourissons ; une enquête de population générale peut être combinée avec un plan de sondage adaptatif ou avec une liste de naissances. Dans de nombreux cas, les enquêtes à bases de sondage multiples peuvent fournir des estimations plus précises des quantités de population sans accroître les coûts de collecte des données, mais le plan et l'estimateur doivent être choisis prudemment afin de réaliser ces économies. Une enquête à

Les enquêtes à bases de sondage multiples peuvent aussi être utilisées conjuguées à des méthodes d'échantillonnage séquentiel ou adaptatif afin d'améliorer le rendement pour une population rare ou difficile à joindre tels que les nouveaux immigrants. Par exemple, un plan de sondage stratifié à plusieurs degrés pourrait être employé pour la base de sondage A, tandis qu'un plan de sondage en grappes adaptatif (Thompson 2002) pourrait être utilisé pour la base de sondage B. Les estimations de domaine peuvent être calculées séparément pour les deux plans de sondage, puis combinées en utilisant les méthodes de la section 2. Dans cette situation, les bases de sondage A et B peuvent se chevaucher entièrement, de sorte que l'erreur de classification dans les domaines ne sera pas un problème.

Tableau 4 $\sqrt{E}M$ estimée pour l'erreur de classification sous deux bases de sondage, avec $n_A = 200$, $n_B = 100$, un échantillon en grappes tiré de la base de sondage A et un échantillon aléatoire simple tiré de la base de sondage B. PEA et PEB désignent les profils d'erreur de classification pour l'erreur de classification dans une enquête à trois bases de sondage, avec $n_A = n_B = n_C = 200$ et un échantillon aléatoire simple tiré de chaque base de sondage. PEA désigne le profil d'erreur de classification pour la base de sondage

PEA	PEB	$\hat{Y}(1/2)$	$\hat{Y}(1/2)_{\text{post1}}$	$\hat{Y}(1/2)_{\text{post2}}$	$\hat{Y}(2/3)$	$\hat{Y}(2/3)_{bc}$	$\hat{Y}(1)$	\hat{Y}_H	\hat{Y}_{FB}	\hat{Y}_{PMV}	\hat{Y}_{PVE}	$\hat{Y}_{BU, \text{rat}}$
a	a	10 916	8 912	8 899	10 916	11 092	11 092	11 879	10 975	9 250	9 155	10 109
a	b	11 786	10 186	10 324	11 157	12 743	11 503	15 463	12 253	8 906	9 391	10 123
a	c	11 983	9 575	9 537	11 409	12 922	11 814	15 600	12 395	9 575	9 279	10 391
a	d	11 042	12 357	12 375	10 941	11 250	11 173	12 056	11 051	11 591	11 605	12 229
b	a	10 698	9 133	9 154	10 872	10 921	11 049	11 875	10 823	9 255	9 151	10 195
b	b	10 957	9 803	9 867	11 071	12 033	11 413	15 262	11 215	8 681	8 748	9 610
b	c	11 115	9 860	9 846	11 272	12 172	11 675	15 361	11 306	9 721	9 252	10 558
b	d	10 988	13 269	13 408	11 046	11 222	11 262	12 143	11 084	12 484	13 347	13 279
c	a	10 995	9 090	9 073	11 106	11 187	11 254	12 028	11 125	9 309	9 190	9 798
c	b	11 104	10 779	11 015	11 090	12 162	11 380	15 348	11 430	9 450	9 724	9 754
c	c	11 155	9 425	9 400	11 189	12 234	11 600	15 424	11 389	9 219	9 064	9 868
c	d	10 922	11 328	11 421	10 896	11 121	11 091	11 929	11 017	10 759	10 456	11 151
d	a	11 011	9 080	9 045	10 920	11 181	11 103	11 913	11 041	9 873	9 579	10 375
d	b	11 838	11 357	11 669	11 164	12 723	11 453	15 299	12 337	10 258	10 848	11 009
d	c	11 804	9 334	9 371	11 159	12 707	11 548	15 298	12 224	9 349	9 507	10 102
d	d	11 179	10 839	10 854	10 989	11 355	11 199	12 059	11 195	10 440	10 302	10 916

Tableau 5

Biais estimé et $\sqrt{E}M$ pour l'erreur de classification dans une enquête à trois bases de sondage, avec $n_A = n_B = n_C = 200$ et un échantillon aléatoire simple tiré de chaque base de sondage. PEA désigne le profil d'erreur de classification pour la base de sondage

1 ; (c) $\phi_{aa}^A = 1, \phi_{ab,ab}^A = 0,9, \phi_{ac,ac}^A = 1, \phi_{abc,abc}^A = 1$; (d) $\phi_{aa}^A = 1, \phi_{ab,ab}^A = 0,9, \phi_{ac,ac}^A = 1, \phi_{abc,abc}^A = 1$; (e) $\phi_{aa}^A = 1, \phi_{ab,ab}^A = 0,8, \phi_{ac,ac}^A = 0,1, \phi_{abc,abc}^A = 1$

PEA	\hat{Y}_{ave}	$\hat{Y}_{ave, post1}$	$\hat{Y}_{ave, post2}$	$\hat{Y}_{ave, bc}$	$\hat{Y}_{2/3}$	$\hat{Y}_{2/3, bc}$	\hat{Y}_{scr}	\hat{Y}_H	\hat{Y}_{PMV}	$\hat{Y}_{BU, \text{rat}}$
a	-8	31	28	-8	5	5	20	9	-26	-208
b	-938	-1 409	-1 478	57	-586	77	107	-2 039	-5 676	-5 624
c	-26	-485	-508	-26	6	6	6	-324	-825	-957
d	-231	-514	-557	104	108	108	85	-326	-1 321	-1 438
e	704	287	247	34	697	27	-4	1 488	1 420	1 193
a	9 003	4 419	4 410	9 003	10 013	10 013	13 108	7 990	7 281	7 293
b	8 961	4 711	4 730	8 955	9 952	9 953	13 092	8 085	9 107	9 074
c	9 119	4 432	4 422	9 119	10 140	10 140	13 238	8 112	7 396	7 422
d	8 894	4 405	4 405	8 893	9 874	9 874	12 919	7 957	7 414	7 433
e	9 088	4 438	4 424	9 059	10 071	10 046	13 180	8 254	7 621	7 581

7. Problèmes de plan de sondage

Comme il est mentionné à la section 1, les plans à bases de sondage multiples peuvent donner une meilleure couverture et une meilleure précision qu'une enquête à base de sondage unique de coût équivalent. Toutefois, le problème d'élaboration du plan de sondage est plus complexe que dans le cas d'une enquête à base de sondage unique, puisqu'un plan de sondage qui est optimal pour la base de sondage A et pour la base de sondage B séparément pourraient ne pas l'être pour l'échantillon combiné. De même, un plan de sondage qui est optimal quand l'estimateur $\hat{Y}(1/2)$ est utilisé peut ne pas l'être pour \hat{Y}_{PMV} .

Hartley (1962, 1974) a obtenu des plans de sondage optimaux pour l'estimateur $\hat{Y}(\theta^H)$ quand un échantillon aléatoire simple est tiré de chaque base de sondage. Les tailles optimales d'échantillons n_A et n_B dépendent des coûts relatifs de l'échantillonnage dans les deux bases de sondages ainsi que des moyennes et des variances de la variable de réponse à l'intérieur des domaines. Cochran (1977, pages 144-145) a décrit l'enquête à base de sondage double de la figure 1 dans son chapitre sur l'échantillonnage stratifié. Dans cette situation, N_a et N_{ab} peuvent être connus, surtout si la base de sondage B est une liste. Les domaines a et ab sont traités comme des strates ; il existe un échantillon provenant de la strate a et deux échantillons

Comme dans l'étude avec deux bases de sondage, les estimateurs corrigés du biais sont approximativement sans biais. L'estimateur avec sélection est également approximativement sans biais puisque $S(A)$ seulement est classifié incorrectement. Les autres estimateurs présentent tous un biais important avec au moins certains des profils d'erreur de classification. Pour les conditions de simulation du tableau 5, les estimateurs poststratifiés, à base de sondage unique (BU) avec ratisage, de Hartley et PMV présentent un grand biais mais ont néanmoins une erreur quadratique moyenne plus petite que les estimateurs à poids fixes et corrigés du biais ; ce classement en fonction de l'EQM ne tient pas dans certaines des autres conditions de simulation. Mecatti (2007) ainsi que Rao et Wu (2010) soutiennent que l'estimateur à poids fixes fondé sur la multiplicité \bar{Y}_{ave} est sans biais si la seule erreur de classification est une erreur entre les domaines qui appartiennent au même nombre de bases de sondage. Classer incorrectement les observations du domaine ab dans le domaine ac (profil c) ne produit pas de biais, parce que l'ajustement des poids dans les deux domaines est $1/2$. Toutefois, en pratique, on s'attendrait à ce que le profil (c), avec deux erreurs d'appartenance à un domaine (ne pas déclarer l'appartenance à la base de sondage B et déclarer correctement l'appartenance à la base de sondage C), soit moins susceptible de survenir en pratique que l'erreur de classification d'une observation appartenant à ab comme appartenant à a ou à abc ; \bar{Y}_{ave} peut être très sensible aux dernières formes d'erreur de classification. Bien qu'un estimateur à poids fixes ne soit pas sensible à l'erreur de classification dans les domaines

Les tableaux 1 à 5 montrent que chaque estimateur décrit à la section 2 peut présenter un biais grave dû à l'erreur de classification dans les domaines. Nous recommandons d'étudier la portée éventuelle de l'erreur de classification entre les domaines durant la phase de pré-test de l'enquête afin que cette information puisse être utilisée dans le plan de sondage. Si les probabilités d'erreur de classification sont connues avec précision, il pourrait être possible de choisir un estimateur à poids fixes qui n'est pas sensible à la forme présomue de l'erreur de classification. Quand on ne peut pas trouver d'estimateur robuste à l'erreur de classification ou quand cet estimateur est inefficace, les estimateurs à poids fixes peuvent être ajustés afin de réduire le biais. Il convient de souligner que les poids corrigés du biais proposés à la section 6.1 sont sensibles aux probabilités d'erreur de classification entrées. Ils ne tiennent pas compte d'autres erreurs non dues à l'échantillonnage, telle que la non-réponse ; l'application des ajustements des poids pour tenir compte de l'erreur de classification décrit à la section 6.1 suivie par les ajustements des poids pour la non-réponse décrits dans Brick et coll. (2011) peut produire des poids finaux qui ne corrigent ni l'erreur de classification ni la non-réponse. En présence à la fois d'erreur de classification dans les domaines et de non-réponse, des ajustements des poids traitant simultanément les deux problèmes sont nécessaires.

Tableau 3
Biais estimé pour l'erreur de classification sous deux bases de sondage, avec $n_A = 200$, $n_B = 100$, un échantillon en grappes tiré de la base de sondage A et un échantillon aléatoire simple tiré de la base de sondage B. PEA et PEB désignent les profils d'erreur de classification pour les bases de sondage A et B

PEA	PEB	$\hat{Y}(1/2)$	$\hat{Y}(1/2)_{post1}$	$\hat{Y}(1/2)_{post2}$	$\hat{Y}(1/2)_{bc}$	$\hat{Y}(2/3)$	$\hat{Y}(2/3)_{bc}$	$\hat{Y}(1)$	\hat{Y}_H	\hat{Y}_{EB}	\hat{Y}_{PMV}	\hat{Y}_{PVE}	$\hat{Y}_{BU,rai}$
a	a	-148	-142	-139	-148	-155	-155	-170	-312	63	-119	-172	-184
a	b	-5 090	4 199	4 599	-72	-6 774	-84	-10 144	-4 976	1 210	3 615	2 181	1 025
a	c	-5 069	-1 088	-851	-72	-6 759	-96	-10 139	-4 800	-1 994	177	-1 136	-3 216
a	d	-39	-8 379	-8 383	-35	-63	-58	-111	-237	-5 757	-5 909	-5 961	-6 996
b	a	1 168	-1 221	-1 258	-79	768	-63	-32	1 395	-1 690	-1 663	-2 514	-3 170
b	b	-3 716	2 979	3 236	60	-5 784	79	-9 918	-2 815	-86	1 776	-346	-2 087
b	c	-3 704	-2 108	-2 074	73	-5 771	92	-9 905	-2 561	-2 970	-1 410	-3 267	-5 814
b	d	1 317	-9455	-9 610	95	926	123	144	1 609	-7 285	-7 317	-7 938	-9 498
c	a	1 179	1 281	1 304	-66	772	-58	-41	1 486	1 831	1 652	943	840
c	b	-3 879	5 545	6 087	-118	-5 971	-126	-10 156	-2 972	3 532	4 597	2 405	1 683
c	c	-3 811	318	636	-44	-5 893	-42	-10 058	-2 671	110	1 128	-784	-2 328
c	d	1 423	-6 858	-6 973	191	1 022	206	220	1 824	-4 328	-4 014	-4 516	-5 624
d	a	-33	2 282	2 290	-28	-35	-32	-40	-148	3 627	3 138	3 103	3 728
d	b	-4 974	6 514	7 123	46	-6 660	30	-10 033	-4 863	4 768	6 274	4 742	4 549
d	c	-4 951	1 412	1 883	80	-6 621	84	-9 961	-4 682	1 357	2 863	1 451	388
d	d	42	-5 987	-5 991	53	40	52	37	-126	-2 899	-2 780	-2 791	-3 317

résultats pour les autres conditions ont produit un profil simulateur et ne sont pas présentés ici. L'estimateur fondé sur la multiplicité \hat{Y}_{avc} , avec $m_i = 1$ pour $i \in \{a, b, c\}$, $m_i = 1/2$ pour $i \in \{ab, ac, bc\}$, et $m_i = 1/3$ pour $i \in abc$, est optimal quand il n'y a pas d'erreur de classification et il est égal à l'estimateur pour base de

sondage unique non ratissé. Les autres estimateurs à poids fixes étudiés sont $\hat{Y}_{2/3}$ avec $m^{(A,a)} = m^{(B,b)} = m^{(C,c)} = 1$, $m^{(A,ab)} = m^{(A,ac)} = m^{(B,ab)} = m^{(B,ac)} = m^{(C,ab)} = m^{(C,ac)} = 1/3$, et $m^{(B,abc)} = m^{(C,abc)} = 1/6$, et l'estimateur avec sélection \hat{Y}_{sct} avec $m^{(A,abc)} = m^{(B,bc)} = m^{(C,c)} = 1$.

Tableau 1
Biais estimé pour l'erreur de classification sous deux bases de sondage, avec $n_A = n_B = 100$ et un échantillon aléatoire simple tiré de chaque base de sondage. PEA et PEB désignent les profils d'erreur de classification pour les bases de sondage A et B

PEA	PEB	$\hat{Y}(1/2)$	$\hat{Y}(1/2)_{post1}$	$\hat{Y}(1/2)_{post2}$	$\hat{Y}(1/2)_{bc}$	$\hat{Y}(2/3)$	$\hat{Y}(2/3)_{bc}$	$\hat{Y}(1)$	\hat{Y}_H	\hat{Y}_{FB}	\hat{Y}_{PMV}	\hat{Y}_{PVE}	$\hat{Y}_{BU, rat}$
a	a	-194	-87	-87	-194	-215	-215	-215	-258	-68	10	-121	-119
a	b	-5 015	4 145	4 529	5	-6 678	17	-10 002	-5 417	1 248	2 486	1 542	2 361
a	c	-5 142	-1 118	-898	-128	-6 823	-138	-10 185	-5 413	-2 583	-1 650	-2 482	-1 690
a	d	-57	-8 430	-8 431	-47	-69	-55	-92	30	-6 576	-6 723	-6 725	-6 795
b	a	1 163	-1 238	-1 290	-82	748	-82	-82	1 355	-2 376	-2 631	-2 551	-2 704
b	b	-3 724	3 040	3 264	43	-5 784	65	-9 905	-3 967	-920	-30	-850	-100
b	c	-3 882	-2 192	-2 187	-124	-5 977	-136	-10 167	-3 954	-4 319	-3 821	-4 477	-3 853
b	d	1 322	-9 445	-9 621	92	917	104	1 600	-8 219	-8 720	-8 531	-8 879	-8 879
c	a	1 366	1 315	1 312	123	969	140	1 530	1 529	1 325	1 355	1 276	1 276
c	b	-3 729	5 456	5 948	51	-5 801	64	-9 945	-4 216	2 096	3 500	2 391	3 355
c	c	-3 797	235	512	-15	-5 868	2	-10 011	-4 089	-1 377	-417	-1 318	-466
c	d	1 285	-7 072	-7 212	56	873	60	48	1 535	-4 665	-5 131	-4 976	-5 222
d	a	-120	2 134	2 134	-111	-132	-126	-155	32	3 710	3 535	3 538	3 470
d	b	-4 979	6 497	7 086	34	-6 620	65	-9 901	-5 599	4 339	5 928	4 788	5 697
d	c	-5 152	1 174	1 644	-137	-6 835	-152	-10 200	-5 622	310	1 626	578	1 540
d	d	90	-5 999	-5 998	107	98	119	114	193	-2 964	-3 116	-3 120	-3 155

Tableau 2
EQM estimée pour l'erreur de classification sous deux bases de sondage, avec $n_A = n_B = 100$ et un échantillon aléatoire simple tiré de chaque base de sondage. PEA et PEB désignent les profils d'erreur de classification pour les bases de sondage A et B

PEA	PEB	$\hat{Y}(1/2)$	$\hat{Y}(1/2)_{post1}$	$\hat{Y}(1/2)_{post2}$	$\hat{Y}(1/2)_{bc}$	$\hat{Y}(2/3)$	$\hat{Y}(2/3)_{bc}$	$\hat{Y}(1)$	\hat{Y}_H	\hat{Y}_{FB}	\hat{Y}_{PMV}	\hat{Y}_{PVE}	$\hat{Y}_{BU, rat}$
a	a	9 646	7 917	7 910	9 646	9 729	9 729	10 304	9 677	8 151	8 081	8 115	8 075
a	b	10 602	9 351	9 531	9 926	11 531	10 197	14 181	11 157	8 212	8 377	8 198	8 311
a	c	10 779	8 622	8 603	10 071	11 715	10 402	14 376	11 243	8 817	8 514	8 720	8 508
a	d	9 789	11 719	11 704	9 674	9 884	9 795	10 432	9 819	10 979	10 978	11 003	11 007
b	a	9 623	8 182	8 185	9 718	9 686	9 766	10 307	9 780	8 446	8 447	8 444	8 459
b	b	9 955	9 054	9 137	9 995	10 949	10 212	14 069	10 489	8 074	7 913	7 995	7 898
b	c	10 146	9 014	9 014	10 160	11 197	10 489	14 404	10 616	9 443	9 108	9 448	9 114
b	d	9 868	12 600	12 716	9 826	9 952	9 927	10 567	10 023	12 063	12 284	12 188	12 371
c	a	9 843	8 185	8 180	9 887	9 853	9 877	10 341	9 991	8 516	8 417	8 442	8 402
c	b	10 049	10 113	10 396	10 039	11 029	10 229	14 127	10 662	8 520	8 863	8 529	8 778
c	c	10 247	8 701	8 718	10 254	11 233	10 534	14 306	10 799	8 762	8 527	8 669	8 516
c	d	10 021	10 861	10 936	9 966	10 068	10 016	10 579	10 177	10 113	10 211	10 168	10 240
d	a	9 795	8 127	8 121	9 734	9 845	9 788	10 343	9 829	9 158	9 024	9 042	8 991
d	b	10 718	10 601	10 970	10 001	11 602	10 258	14 149	11 358	9 461	10 157	9 595	9 986
d	c	10 847	8 558	8 650	10 099	11 769	10 426	14 387	11 424	8 674	8 707	8 608	8 664
d	d	9 945	10 070	10 057	9 778	10 019	9 885	10 510	9 986	9 458	9 412	9 449	9 417

variance obtenue par poststratification est supérieure ou non au biais supplémentaire qui peut être introduit. Le calage sur les totaux des bases de sondage N_A et N_B , dans $X_{\text{SF, rat.}}$, a le même effet que la poststratification sur le biais d'erreur de classification.

Pour les échantillons aléatoires simples des tableaux 1 et 2, les estimateurs PMV et PVE présentent souvent un nettement plus grand biais que les estimateurs à poids fixes non corrigés. Pour ces méthodes, les contributions relatives des deux bases de sondage dépendent des variances estimées de N_A^{ab} et N_B^{ab} , et ces deux facteurs interagissent de façon complexe selon la structure de l'erreur de classification. Pour le profil d'erreur de classification (d) dans l'une ou l'autre base de sondage, N_{PMV}^{ab} est trop petit parce que les observations dans le domaine ab sont classées incorrectement; par conséquent, les poids des observations comprises dans les domaines non chevauchants sont trop grands. Une version poststratifiée de l'estimateur PMV présente les problèmes de biais des estimateurs à poids fixes poststratifiés. En forçant les estimateurs de Y_{ab}^{ab} à être égaux, l'estimateur PVE peut aggraver le biais. Par exemple, dans la situation de la ligne 3 du tableau 1, avec une classification correcte pour la base de sondage A et le profil (c) pour la base de sondage B, le biais de PVE est supérieur de 50 % au biais de PMV. Dans ce cas, l'estimateur PVE tire l'estimateur biaisé Y_{ab}^{ab} provenant de $S(A)$ vers l'estimateur biaisé provenant de la base de sondage B. Les estimateurs opti-

maux ont aussi un biais important.

Quand un échantillon en grappes est tiré de la base de sondage A comme aux tableaux 3 et 4, les profils de biais sont similaires. Quand il n'y a pas d'erreur de classification, les EQM des estimateurs optimaux et PMV sont plus faibles que celles de $Y(2/3)$, parce qu'elles tiennent compte du plan de sondage. Toutefois, en présence d'erreur de classification, l'avantage en ce qui concerne l'EQM est réduit à cause de l'accroissement du biais.

Pour étudier l'erreur de classification dans une enquête à trois bases de sondage, nous avons tiré des échantillons aléatoires simples de chaque base de sondage et nous avons les classifications correctes pour les bases de sondage B et C. Le tableau 5 donne les résultats pour une simulation avec trois bases de sondage et un échantillon aléatoire simple de taille 200 provenant de chacune des bases. La population a été générée avec $N_a = 10\,000$ dans chaque domaine et les moyennes de domaine $\mu_a = 1$, $\mu_{ab} = 2$, $\mu_{ac} = 3$, $\mu_{abc} = 4$, $\mu_b = 5$, $\mu_{bc} = 6$, $\mu_c = 7$. Dans cette simulation, les bases de sondages B et C sont classées correctement et les profils d'erreur de classification pour la base de sondage A sont donnés dans le tableau. Nous avons aussi étudié d'autres moyennes de domaine, tailles de domaine de population et tailles d'échantillon en nous servant d'un plan factoriel; les

dans a et $m_i^a = 1/2$ pour i classé dans ab ; pour le profil (c), les ajustements des poids corrigés du biais sont 17/16 et 7/16, respectivement. L'estimateur pour base de sondage unique est omis de ces tableaux puisqu'il est le même que $Y(1/2)$ ou $Y(2/3)$; l'estimateur pour base de sondage unique calé par ratisage sur les totaux de population N_A et N_B est désigné par $X_{\text{SF, rat.}}$. Les tableaux 1 et 2 donnent les résultats pour $n_A = 100$, $n_B = 100$, $N_a = N_b = 25\,000$, et un échantillon en grappes tiré de la base de sondage A. Les profils généraux des résultats sont similaires pour les autres simulations qui ne sont pas présentées ici.

Premièrement, considérons les estimateurs à poids fixes. Les estimateurs corrigés du biais réduisent le biais comme prévu; dans tous les cas étudiés avec erreur de classification, le biais empirique provenant des estimateurs corrigés du biais étaient inférieurs à 200 en valeur absolue, ce qui est compris dans la marge d'erreur. Bien que l'écart-type pour les estimateurs corrigés du biais soit plus élevé que pour les estimateurs non corrigés, dans la plupart des cas, les erreurs quadratiques moyennes sont comparables.

L'estimateur avec sélection $Y(1)$, qui élimine les unités provenant de la base de sondage B du domaine ab , ne présente pas de biais d'erreur de classification quand les unités de la base de sondage B sont classées correctement. Il ne présente pas non plus de biais dans les tableaux 1 et 3 dans le cas du profil (d) d'erreur de classification dans la base de sondage B, parce que les observations classées incorrectement dans le domaine b au lieu du domaine ab ont une moyenne nulle; pour différents ensembles de moyennes de domaine, le profil (d) ne crée pas de biais. Pour les autres cas, l'estimateur avec sélection est celui dont le biais est le plus important. Pour chaque profil d'erreur de classification, l'estimateur avec sélection possède une plus grande erreur quadratique moyenne parce que des données sont écartées. Si les moyennes de domaine sont similaires, l'erreur de classification dans les domaines pourrait ne pas causer de biais appréciable, mais le rejet des observations provenant du domaine ab dans $S(B)$ augmenterait fortement l'erreur quadratique moyenne.

La poststratification sur les totaux de domaine en présence d'erreurs de classification augmente souvent le biais au lieu de le réduire. Considérons la ligne 4 du tableau 1, où 20 % des observations de $S(B)$ appartenant au domaine ab sont classées par erreur dans le domaine b . Les poids des observations qui appartiennent réellement au domaine b , de moyenne 2, sont réduits de 500 à environ 417, ce qui cause un biais dans les versions poststratifiées de $Y(1/2)$. L'effet de la poststratification sur l'erreur quadratique moyenne est mixte et dépend du fait que la réduction de la

multinomial, mais augmente la variance. Partant de la

base de sondage A_q

$$V \left[\sum_{g=1}^G (\mathbf{m}_{A_q}^g)' \mathbf{Y}_{A_q}^g (\text{mis}, g) \right]$$

$$= E \left[V \left(\sum_{g=1}^G (\Phi_{A_q}^g)' (\mathbf{m}_{A_q}^g)' (\mathbf{M}_{A_q}^g)' \delta_{A_q}^g \chi_i(g) w_{A_q}^g y_i \mid S(A_1), \dots, S(A_{\bar{q}}) \right) \right]$$

$$+ V \left[E \left(\sum_{g=1}^G (\Phi_{A_q}^g)' (\mathbf{m}_{A_q}^g)' (\mathbf{M}_{A_q}^g)' \delta_{A_q}^g \chi_i(g) w_{A_q}^g y_i \mid S(A_1), \dots, S(A_{\bar{q}}) \right) \right]$$

$$\delta_{A_q}^g \chi_i(g) w_{A_q}^g y_i \mid S(A_1), \dots, S(A_{\bar{q}}) \Big]$$

$$= \sum_G [(\Phi_{A_q}^g)' (\mathbf{m}_{A_q}^g)' (\mathbf{M}_{A_q}^g)' \chi_i(g) (w_{A_q}^g y_i)^2$$

$$\{ \text{diag}[(\Phi_{A_q}^g)' \delta_{A_q}^g] - (\Phi_{A_q}^g)' \delta_{A_q}^g (\delta_{A_q}^g)' \Phi_{A_q}^g \} (\Phi_{A_q}^g)' (\mathbf{M}_{A_q}^g)' + \mathbf{m}_{A_q}^g \left[\sum_{i \in S(A_q)} \chi_i(g) w_{A_q}^g y_i \right] + V \left[\sum_{i \in S(A_q)} \{ \mathbf{m}_{A_q}^g \}' \delta_{A_q}^g w_{A_q}^g y_i \right].$$

Le deuxième terme est la variance de la contribution de la base de sondage A_q quand les unités sont classées correctement. Le premier terme est nul uniquement si $\Phi_{A_q}^g$ est diagonale pour tout g , c'est-à-dire qu'il n'y a pas d'erreur de classification.

Les ajustements des poids donnés par (6) peuvent être étendus au cas dans lequel les poids fixes originaux \mathbf{m}^{A_q} varient pour les groupes, à condition que $\sum_{g=1}^G \mathbf{m}_{(A_q, g)}^g = \mathbf{1}$ pour chaque domaine. Notons que la méthode de correction du biais exposée à la présente section est proposée uniquement pour les estimateurs à poids fixes et non pour les estimateurs PMV, PVE ou optimaux dans lesquels les poids fondés sur la multiplicité dépendent des données. La correction du biais dépend de la spécification correcte des probabilités d'erreur de classification. Si ces dernières sont estimées d'après une autre enquête, les méthodes d'opérations des enquêtes doivent être semblables.

6.2 Étude en simulation

Lohr et Rao (2006) ont constaté dans des études en simulation que l'estimateur PMV avait une plus petite erreur quadratique moyenne que les autres estimateurs en présence d'erreur de classification aléatoire, mais cela tient en grande partie à la variance plus faible que cet estimateur. Pour étudier la sensibilité des estimateurs à d'autres formes d'erreur de classification dans les domaines, nous avons exécuté une étude en simulation pour des enquêtes à deux et à trois bases de sondage. La population du domaine d a été générée en utilisant le modèle $y_{ij} = \mu_d + \alpha_i + \varepsilon_{ij}$ pour $i = 1, \dots, N_d$ et $j = 1, \dots, 5$, avec $\alpha_i \sim N(0, 1)$ et

$\varepsilon_{ij} \sim N(0, 1)$ générée indépendamment, puis des échantillons probabilistes ont été tirés de cette population. Pour l'étude à deux bases de sondage, les moyennes de domaine sont $\mu_a = -1$, $\mu_{ab} = 0$, $\mu_b = 2$ et les facteurs de la simulation sont :

1. Taille de l'échantillon : 100 ou 200 provenant de

chaque base de sondage.

2. Échantillon en grappes ou échantillon aléatoire simple tiré de la base de sondage A. Un échantillon en grappes a été tiré en sélectionnant un échantillon aléatoire simple de $n_4/5$ des groupes utilisés pour générer la population.

3. Probabilités d'erreur de classification pour la base de sondage A (toutes les probabilités non énumérées sont égales à 0) :

a. $\phi_{Aa}^{aa} = 1$, $\phi_{Aa,ab}^{aa} = 1$ (pas d'erreur de classification) ;

b. $\phi_{Aa}^{aa} = 0,9$, $\phi_{Aa,ab}^{aa} = 0,1$, $\phi_{Aa,ab}^{aa,ab} = 1$;

c. $\phi_{Aa}^{aa} = 0,9$, $\phi_{Aa,ab}^{aa} = 0,1$, $\phi_{Aa,ab}^{aa,ab} = 0,1$, $\phi_{Aa,ab}^{aa,ab,ab} = 0,1$;

d. $\phi_{Aa}^{aa} = 1$, $\phi_{Aa,ab}^{aa} = 0,9$, $\phi_{Aa,ab}^{aa,ab} = 0,1$.

4. Probabilités d'erreur de classification pour la base de sondage B :

a. $\phi_B^{bb} = 1$, $\phi_{B,ab}^{bb} = 1$ (pas d'erreur de classification) ;

b. $\phi_B^{bb} = 0,8$, $\phi_{B,ab}^{bb} = 0,2$, $\phi_{B,ab}^{bb,ab} = 1$;

c. $\phi_B^{bb} = 0,8$, $\phi_{B,ab}^{bb} = 0,2$, $\phi_{B,ab}^{bb,ab} = 0,9$, $\phi_{B,ab}^{bb,ab,ab} = 0,1$;

d. $\phi_B^{hh} = 1$, $\phi_{B,ab}^{hh,ab} = 0,8$, $\phi_{B,ab}^{hh,ab,ab} = 0,2$.

5. Tailles de population : $N_a = N_b = N_{ab} = 25\,000$; $N_a = N_b = 10\,000$, $N_{ab} = 55\,000$; $N_a = 40\,000$, $N_b = 10\,000$.

Dix mille répliques ont été exécutées pour chaque combinaison de facteurs, ce qui donne à l'estimation Monte Carlo du biais une erreur-type d'environ 100. Nous avons étudié tous les estimateurs décrits à la section 2, y compris $Y(1/2)$, $Y(2/3)$, et $Y(1)$ provenant de l'équation (3). Nous avons également examiné les estimateurs poststratifiés qui pourraient être employés quand les chiffres de population de domaine N_a , N_{ab} et N_b sont connus : les estimateurs avec l'indice inférieur « post1 » appliquent la poststratification aux deux échantillons pour commencer puis combinent les échantillons, tandis que les estimateurs avec l'indice inférieur « post2 » combinent les échantillons d'abord, puis font la poststratification sur les chiffres de population de domaine. Les estimateurs corrigés du biais $Y(1/2)_h$ et $Y(2/3)_h$ modifient les poids fixes initiaux correspondant à $\theta = 1/2$ et $\theta = 2/3$ en utilisant (6). Avec le profil d'erreur de classification (b) dans la base de sondage A, par exemple, les ajustements des poids corrigés du biais pour $Y(1/2)^{bc}$ sont $m_i^1 = 19/18$ pour i classé

base de sondage. Dans une enquête téléphone fixe/téléphone mobile, on pourrait savoir que la probabilité d'une erreur de classification est plus élevée pour certains groupes d'âge que pour d'autres. Chambers, Chipperfield, Davis et Kováčević (2008) ont utilisé une approche de groupement similaire pour corriger les erreurs de couplage d'enregistrement. Supposons que la population peut être divisée en G groupes, $g = 1, \dots, G$, dans lesquels les probabilités d'erreur de classification sont connues pour chaque base de sondage A_q^g . Soit $\phi_{A_q^g}^g(d, e)$ la probabilité qu'une observation dans le groupe g avec domaine vrai d soit classée dans le domaine e dans l'échantillon $S(A_q^g)$, et soit $\Phi_{A_q^g}^g$ la matrice de dimension $D \times D$ dont les entrées sont $\phi_{A_q^g}^g(d, e)$. Pour l'observation i appartenant au groupe g et au vrai domaine d , supposons que la ligne d de $\mathbf{M}_{A_q^g}^i$ est générée sous forme d'une variable aléatoire multinominale de taille 1 avec les probabilités dans la ligne d de la matrice des erreurs de classification attendues $\Phi_{A_q^g}^g$, et que tous les $\mathbf{M}_{A_q^g}^i$ sont indépendants les uns des autres et des variables d'inclusion dans l'échantillon. Nous avons donc G matrices de probabilités d'erreur de classification pour la base de sondage A_q^g , $\Phi_{A_q^g}^1, \dots, \Phi_{A_q^g}^G$. Désignons le vecteur de totaux de population pour le groupe g par $\mathbf{Y}(g) = \sum_{i=1}^N \delta_{A_q^g}^i \chi_i(g) Y_i$, où $\chi_i(g) = 1$ si l'observation i est dans le groupe g et 0 autrement.

Pour les classifications dans les domaines observées A_q^g , l'estimateur pondéré par les poids de sondage du vecteur des totaux de population dans le groupe g est

$$\mathbf{Y}_{A_q^g}^g(\text{mis}, g) = \sum_{i \in S(A_q^g)} \mathbf{\eta}_{A_q^g}^i \chi_i(g) w_{A_q^g}^i Y_i$$

$$= \sum_{i \in S(A_q^g)} (\mathbf{M}_{A_q^g}^i)' \delta_{A_q^g}^i \chi_i(g) w_{A_q^g}^i Y_i$$

de sorte que $E[\mathbf{Y}_{A_q^g}^g(\text{mis}, g)] = (\Phi_{A_q^g}^g)' \mathbf{Y}(g)$.

Considérons maintenant un nouveau vecteur d'ajustements des poids $\tilde{\mathbf{m}}_{A_q^g}^g = (\tilde{m}_{A_q^g, 1}^g, \dots, \tilde{m}_{A_q^g, D}^g)'$ pour le groupe g dans la base de sondage A_q^g . Alors

$$E \left[\sum_{\tilde{O}} \sum_{g=1}^G (\tilde{\mathbf{m}}_{A_q^g}^g)' \mathbf{Y}_{A_q^g}^g(\text{mis}, g) \right] = \sum_{\tilde{O}} \sum_{g=1}^G (\Phi_{A_q^g}^g \tilde{\mathbf{m}}_{A_q^g}^g)' \mathbf{Y}(g).$$

Puisque $\sum_{\tilde{O}} \sum_{g=1}^G (\mathbf{m}_{A_q^g}^g)' \mathbf{Y}(g) = Y$, le biais sera éliminé sous ce modèle quand

$$\tilde{\mathbf{m}}_{A_q^g}^g = (\Phi_{A_q^g}^g)' \mathbf{m}_{A_q^g}^g, \quad (6)$$

où $(\Phi_{A_q^g}^g)^+$ est l'inverse de Moore-Penrose de $\Phi_{A_q^g}^g$, obtenue en prenant l'inverse des lignes et des colonnes non nulles de $\Phi_{A_q^g}^g$. Le remplacement des ajustements $\mathbf{m}_{A_q^g}^g$ par $\tilde{\mathbf{m}}_{A_q^g}^g$ élimine le biais sous le modèle d'erreur de classification

En pratique, nous attendons à ce que les erreurs de classification dans les domaines soient reliées aux réponses d'intérêt ; nous nous attendons aussi à ce que, dans de nombreuses situations, l'erreur de classification soit plus probable dans certaines directions. Dans les enquêtes longitudinales à deux bases de sondage, l'erreur de classification dans les domaines peut avoir des effets plus importants que dans les enquêtes transversales (Lu et Lohr 2010). Dans certaines situations, l'indicateur de domaine peut manquer ou ne pas être disponible. Clark, Winglee et Liu (2007) ont étudié des méthodes de régression logistique et des méthodes d'appariement d'enregistrements pour prédire le domaine d'une observation pour laquelle manque l'information sur le domaine.

6.1 Corrections du biais dû à l'erreur de classification

Si l'erreur de classification dans les domaines est importante, chaque méthode de modification des poids de sondage en vue de les ajuster pour tenir compte de la multiplicité peut produire des estimations biaisées des quantités de population. À la présente section, nous étalibsons une correction pour le biais dû à l'erreur de classification dans les domaines de l'estimateur à poids fixes de la section 2.2 quand les probabilités d'erreur de classification sont connues. Soit le D -vecteur $\delta_{A_q^g}^i$ qui désigne l'appartenance vraie au domaine pour l'observation i de la base de sondage A_q^g , contenant un 1 en position d si l'observation i est dans le domaine d , et 0 autrement. Soit $\mathbf{Y} = (Y_1, \dots, Y_D)'$ le vecteur des totaux de population pour les D domaines. Pour une enquête à deux bases de sondage chevauchantes, $\mathbf{Y} = (Y^a, Y^{ab}, Y^b)'$, pour une enquête à trois bases de sondage, $\mathbf{Y} = (Y^a, Y^{ab}, Y^{ac}, Y^{abc}, Y^b, Y^c)'$. S'il n'y a pas d'erreur de classification dans les domaines,

$$\mathbf{Y}_{A_q^g}^g = \sum_{i \in S(A_q^g)} \delta_{A_q^g}^i w_{A_q^g}^i Y_i$$

est l'estimateur correspondant de \mathbf{Y} provenant de $S(A_q^g)$. Pour le vecteur d'ajustement des poids fixes $\mathbf{m}_{A_q^g}^g = (m_{A_q^g, 1}^g, \dots, m_{A_q^g, D}^g)'$ dans la base de sondage A_q^g , satisfaisant $\sum_{\tilde{O}} \mathbf{m}_{A_q^g}^g = 1$, alors $E[\sum_{\tilde{O}} (\mathbf{m}_{A_q^g}^g)' \mathbf{Y}_{A_q^g}^g] = Y$.

Maintenant, supposons qu'il existe une erreur de classification. Soit $\mathbf{\eta}_{A_q^g}^i$ la classification observée de l'observation i dans S . Nous pouvons écrire $\mathbf{\eta}_{A_q^g}^i = (\mathbf{M}_{A_q^g}^i)' \delta_{A_q^g}^i$, où $\mathbf{M}_{A_q^g}^i$ est une matrice de dimensions $D \times D$ contenant un 1 en position (d, e) si l'observation i dans le vrai domaine d est classée (in)correctement comme appartenant au domaine e , et 0 autrement.

Pour tenir compte de l'erreur de classification différente à l'intérieur des domaines, nous posons une structure dans laquelle les probabilités d'erreur de classification peuvent différer selon la sous-population dans une

présence d'un effet de mode s'il n'existe pas d'autres erreurs non dues à l'échantillonnage. Sinon, une grande valeur de D_{ab} ne fournit pas d'information sur la cause de la différence : une expérimentation est nécessaire pour faire la distinction entre les causes possibles.

6. Erreur de classification dans les domaines et correction du biais

Dans les estimateurs examinés à la section 2, les poids des observations sont construits en se basant sur l'appartenance à un domaine. Donc, dans l'estimateur $\hat{Y}(\theta)$ donné par (3), le multiplicateur du poids d'une observation provenant de l'échantillon tiré de la base de sondage A est 1 si l'observation est dans le domaine a , et est θ si l'observation est dans le domaine ab , afin de tenir compte de la multiplicité de l'échantillonnage.

En pratique, l'appartenance au domaine n'est pas toujours claire. Dans le cas de la figure 1, on peut ne pas savoir si un répondant dans une base de sondage aréolaire appartient aussi à la liste. Si la base de sondage A est une base aréolaire et que la base de sondage B est une base de sondage en ligne, par exemple, le seul moyen de déterminer si une personne échantillonnée à partir de la base de sondage A figure aussi dans la base de sondage B pourrait consister à demander à la personne si elle a accès à Internet, et la personne pourrait ne pas donner la réponse correcte.

Si l'on recourt à l'appariement ou au couplage d'enregistrements pour déterminer l'appartenance à une base de sondage, un appariement imparfait peut également produire une classification incorrecte des observations. Lesser et Kalsbeek (1999) ont discuté des erreurs non dues à l'échantillonnage qui ont lieu dans les enquêtes à deux bases de sondage menées par le U.S. National Agricultural Statistics Service. Une erreur de classification dans les domaines peut avoir lieu si une ferme échantillonnée dans la base aréolaire est classée incorrectement en ce qui concerne son appartenance à la liste. Dans les enquêtes téléphoniques à deux bases de sondage fixe/téléphone fixe/téléphone, il est difficile de déterminer si une personne présente dans une base de sondage est également présente dans l'autre (Kennedy 2007). Une personne rejointe dans un échantillon de numéros de téléphone fixe peut aussi posséder un téléphone mobile, mais rarement prendre des appels sur le téléphone mobile. Bien qu'elle appartienne techniquement au domaine de chevauchement, cette personne ne peut pour ainsi dire pas être rejointe dans l'enquête utilisant les téléphones mobiles. Dans certaines enquêtes téléphoniques/téléphone mobile, on demande aux répondants d'indiquer les parts relatives d'usage du téléphone mobile ou du téléphone fixe, mais une erreur de classification peut avoir lieu.

Bon nombre d'estimateurs pour enquête à bases de sondage multiples combinent les estimations provenant des domaines de chevauchement, et ces méthodes reposent sur l'hypothèse que les estimateurs de X_{ab}^{ab} provenant des enquêtes composantes estiment tous deux la même quantité. Si, par contre, l'enquête à base de sondage A est menée sur place, tandis que l'enquête à base de sondage B est menée par téléphone, il est possible qu'un recensement du domaine ab d'après la base de sondage B donne un total de domaine différent du recensement d'après la base de sondage A.

Un moyen d'étudier les effets de mode consiste à réaliser l'enquête à base de sondage B en utilisant un échantillon scindé, par exemple interviewé en partie sur place et en partie par téléphone, mais cela réduirait l'économie due à l'utilisation de deux bases de sondage. Un pré-test minutieux peut atténuer les effets de mode. Des travaux de recherche sont nécessaires dans ce domaine ; le même problème d'effets de mode se pose évidemment dans les enquêtes à base de sondage unique, tel que l'American Community Survey, dans laquelle le suivi des cas de non-réponse est effectué selon un mode différent de celui utilisé pour l'échantillon original (voir Citro et Kalton 2007). Les méthodes présentées dans Leeuw, Hox et Dillman (2008) pour la conception des enquêtes dans le cas de modes de collecte multiples s'appliquent aussi au cas des enquêtes à bases de sondage multiples.

Vannieuwenhuysen, Loosveldt et Molenberghs (2011) ont présenté une méthode pour distinguer les effets de mode des effets de sélection quand une enquête à mode unique complètement est disponible. Ils ont toutefois constaté que la méthode requiert l'hypothèse forte selon laquelle les erreurs de couverture et de non-réponse sont équivalentes dans les deux enquêtes. Si cette hypothèse est satisfaite pour une enquête à deux bases de sondage, de sorte que les échantillons dans le domaine de chevauchement provenant des bases de sondage A et B représentent la même population, et que la classification par domaine est correcte, l'effet de mode peut être estimé d'après le domaine de chevauchement comme étant $D_{ab} = \hat{Y}_{ab}^A - \hat{Y}_{ab}^B$. Une différence qui diffère de manière significative de 0 indique la

rééchantillonnage peut être poststratifiée sur les totaux de population et de base de sondage, de sorte que la post-stratification soit prise en compte dans l'estimation de la variance.

L'une des difficultés que posent les méthodes d'estimation de la variance par rééchantillonnage tient au fait qu'un très grand nombre de colonnes de poids de rééchantillonnage peut être nécessaire si un échantillon aléatoire simple ou un échantillon aléatoire stratifié est tiré de l'une des bases de sondage. Dans le cas du bootstrap, nous avons constaté que, pour certaines enquêtes, au moins 500 itérations bootstrap sont requises pour les estimations de la variance dans le cas d'enquêtes à deux bases de sondage, ce qui de nouveau peut être excessif. Il se peut que l'estimation de la variance pour des strates combinées, comme il est discuté dans Lu, Brick et Sitter (2006), soit utilisable dans le cas des enquêtes à bases de sondage multiples en vue de réduire le nombre de répliques nécessaires.

5. Erreurs non dues à l'échantillonnage

Les enquêtes à bases de sondage multiples ont souvent une meilleure couverture de la population que les enquêtes à bases de sondage unique. Quand toutes les bases de sondage sont incomplètes, comme à la figure 3, n'importe laquelle des bases de sondage A, B ou C, si elle était utilisée comme seule base de sondage, donnerait lieu à un sous-dénombrement grave. Les plans à bases de sondage multiples font en sorte que toutes les unités qui se trouvent dans les bases de sondage chevauchantes ont une probabilité d'inclusion positive.

Comme toutes les autres enquêtes, les enquêtes à bases de sondage multiples sont sujettes à des erreurs non dues à l'échantillonnage. Elles souffrent de non-réponse, qui peut être différente dans les diverses bases de sondage. Même si l'union des bases de sondage peut produire une meilleure couverture qu'une base de sondage unique, il pourrait persister un sous-dénombrement de la population cible. Les estimateurs pour les enquêtes à bases de sondage multiples sont également sensibles aux erreurs de classification dans les domaines et aux biais qui pourraient résulter de différentes méthodes ou de différents modes de collecte dans les enquêtes composantes. Nous discutons de la non-réponse et des effets de mode à la présente section, et nous étudions les effets de l'erreur de classification dans les domaines à la section 6.

5.1 Non-réponse

Dans toute enquête, la non-réponse peut introduire un biais dans les estimations des totaux de population et d'autres quantités. Des taux de non-réponse différents dans les échantillons provenant des deux bases de sondage peuvent avoir une incidence sur les estimations ponctuelles

du total de population données à la section 2 ; en outre, la non-réponse peut affecter les ajustements des poids prescrits par certaines méthodes.

Kennedy (2007) a discuté d'un problème qui a été constaté quand la base de sondage A est une liste de numéros de téléphone fixe et que la base de sondage B est une liste de numéros de téléphone mobile : les unités dans le domaine d'intersection *ab* qui ont répondu sur un téléphone mobile diffèrent de celles qui l'ont fait sur un téléphone fixe. Par exemple, on a estimé que 18 % des unités dans le domaine d'intersection avaient de 18 à 25 ans dans l'échantillon tiré de la base de sondage B, tandis que 8 % seulement des unités de l'intersection avaient de 18 à 25 ans en se servant de l'échantillon de la base de sondage A. La différence a été attribuée à la non-réponse : on a pensé que les personnes qui utilisent principalement un téléphone mobile (et donc sont difficiles à joindre dans le cadre d'une enquête fondée sur des numéros de téléphone fixe) ont tendance à être plus jeunes. Kennedy (2007) a suggéré un calage (*raking*) en se servant de l'usage relatif estimé du téléphone (c'est-à-dire si la plupart des appels sont faits sur un téléphone fixe ou sur un téléphone mobile).

Brick et coll. (2011) ont proposé deux méthodes pour l'ajustement de la non-réponse dans les enquêtes téléphoniques à deux bases de sondage (lignes mobiles/lignes fixes) avec des estimateurs à poids fixes. Ils ont considéré les conditions dans lesquelles le domaine de chevauchement possède deux groupes : les ménages qui reçoivent tous leurs appels, ou presque tous, sur des téléphones mobiles (mobile principalement) et les autres ménages dans le domaine de chevauchement (fixe principalement). La première méthode, qui ne requiert pas d'estimation externe des totaux de contrôle, consiste à fixer la valeur de θ dans l'estimateur avec ajustement des poids fixes de manière à réduire le biais de non-réponse en utilisant les taux de réponse pour les ménages se servant d'un téléphone mobile principalement et d'un téléphone fixe principalement dans chaque échantillon. La deuxième méthode nécessite des totaux de contrôle de poststratification pour les groupes à téléphone mobile principalement et téléphone fixe principalement dans le domaine chevauchant, N_{1ab} et N_{2ab} , et estime le total de population dans le domaine *ab* par

$$\sum_{g=1}^G \theta_g \frac{N_{gA}^{gab}}{N_{gab}^{gab}} Y_{gab}^A + (1 - \theta_g) \frac{N_{gB}^{gab}}{N_{gab}^{gab}} Y_{gab}^B,$$

où Y_{gab}^A représente le total estimé du groupe *g* dans le domaine *ab* provenant de $S(A)$, les autres totaux sont définis de la même façon et $0 \leq \theta_g \leq 1$ pour $g = 1, 2$.

5.2 Effets de mode

Dans certains cas, les bases de sondage multiples peuvent aussi signifier modes multiples. De Leeuw (2008)

pour l'estimateur (3), dans lequel les ajustements des poids ne dépendent pas des données. Dans cette situation,

$$V[\hat{Y}(\theta)] = V\left[\sum_{i \in S(A)} w_i^A y_i\right] + V\left[\sum_{i \in S(B)} w_i^B y_i\right],$$

où w_i^A et w_i^B sont définis sous (2). Créer l'ensemble de données par concaténation des observations provenant de $S(A)$ et $S(B)$ comme à la section 4.1, en utilisant w_i^A et w_i^B comme poids. Définir la variable de stratification pour l'échantillon combiné comme étant la combinaison des catégories données par la variable indicatrice de la base de sondage, la variable de stratification de la base de sondage A et la variable de stratification de la base de sondage B. Définir la variable de groupement de premier degré pour l'échantillon combiné de la même façon que pour la combinaison des catégories des variables de groupement des bases de sondage individuelles. Ensuite, un logiciel d'analyse de données d'enquête standard peut être utilisé pour estimer les moyennes et les totaux de population en se servant des poids modifiés, et pour estimer les variances en se servant des variables de stratification et des variables de groupement provenant des échantillons combinés.

L'estimation de la variance est plus compliquée quand les modifications des poids m_i^A ou m_i^B dépendent de quantités qui sont estimées d'après l'échantillon, comme dans l'estimateur PMV, ou quand l'échantillon combiné est poststratifié ou calé sur des quantités de population. Les méthodes de linéarisation, du jackknife et du bootstrap peuvent alors être employées pour estimer les variances.

Voici maintenant un résumé des méthodes qui peuvent être utilisées pour estimer la variance si les UPE provenant des bases de sondage sont sélectionnées indépendamment. Quand les échantillons provenant des diverses bases de sondage ont des UPE en commun, d'autres méthodes doivent être utilisées. Si, par exemple, les UPE sont tirées de la population et qu'un plan à deux bases de sondage est utilisé dans chaque UPE sélectionnée, les estimateurs ponctuels des totaux d'UPE peuvent être calculés en utilisant l'une des méthodes décrites à la section 2. Dans ce cas, les méthodes de rééchantillonnage classiques peuvent être utilisées pour calculer un estimateur de variance avec remplacement.

Sous certaines conditions de régularité, les méthodes de linéarisation et du jackknife sont convergentes pour l'estimation de la variance d'une caractéristique de population τ qui peut s'écrire sous la forme $\tau = g(\mathbf{A}, \mathbf{B})$, où \mathbf{A} est un vecteur de totaux de population provenant de la base de sondage A, \mathbf{B} est un vecteur de totaux de population provenant de la base de sondage B, et g est une fonction de deux fois continuellement dérivable (Skinner et Rao 1996; Lohr et Rao 2000). Le vecteur \mathbf{A} est estimé d'après $S(A)$ par $\hat{\mathbf{A}}$, avec la matrice de covariance estimée $\hat{\Sigma}_A$; de

même, \mathbf{B} estime $S(B)$, avec $V(\mathbf{B}) = \hat{\Sigma}_B$. L'estimateur par linéarisation de la variance de $\hat{\tau} = g(\mathbf{A}, \mathbf{B})$ est

$$V^L(\hat{\tau}) = g_A^L \sum_{i \in A} g_A^L + g_B^L \sum_{i \in B} g_B^L,$$

où g_A^L est le vecteur de dérivées partielles de $g(\mathbf{A}, \mathbf{B})$ par rapport aux composantes de \mathbf{A} et g_B^L est le vecteur correspondant de dérivées partielles provenant de la base de sondage B. Demnati, Rao, Hidiroglou et Tamby (2007) ont obtenu les estimateurs de la variance par linéarisation pour les enquêtes à bases de sondage multiples en prenant les dérivées d'une fonction des poids plutôt que des moyennes. Les méthodes de linéarisation requièrent que les dérivées soient calculées séparément pour chaque estimateur pris en considération, et ces calculs peuvent être fastidieux. Pour cette raison, il est parfois préférable d'utiliser les méthodes de rééchantillonnage quand des enquêtes à bases de sondage multiples sont adoptées.

Supposons que l'on tire un échantillon stratifié à plusieurs degrés comprenant H strates de la base de sondage A, où la strate h contient n_h^A unités primaires d'échantillonnage. Un échantillon stratifié à plusieurs degrés indépendant comprenant L strates est tiré de la base de sondage B, où la strate l contient n_l^B unités primaires d'échantillonnage. L'estimateur jackknife de la variance peut être calculé en créant un total de $\sum_{h=1}^H n_h^A + \sum_{l=1}^L n_l^B$ colonnes de poids de rééchantillonnage (Lohr et Rao 2000). Les poids de rééchantillonnage pour la colonne correspondant à la suppression de l'UPE i de la strate h dans S_A sont donnés par :

$$\tilde{w}_k^A = \begin{cases} \frac{n_h^A}{n_h^A - 1} & \text{si l'unité } k \text{ est dans la strate } h \text{ mais non dans l'UPE } i; \\ 0 & \text{si l'unité } k \text{ est dans l'UPE } i \text{ de la strate } h; \\ w_k^A & \text{si l'unité } k \text{ est dans la strate } g \neq h. \end{cases}$$

Le coefficient jackknife pour cette colonne est le multiplicateur $(n_h^A - 1) / n_h^A$. La colonne de poids de rééchantillonnage correspondant à la suppression de l'UPE j provenant de la strate l dans S_B est formée de la même façon, avec le coefficient jackknife $(n_l^B - 1) / n_l^B$. Lorsqu'il y a plus de deux bases de sondage, des colonnes supplémentaires de poids de rééchantillonnage correspondant aux UPE supprimées provenant de ces échantillons sont ajoutées. Les poids pour une méthode bootstrap d'estimation de la variance (voir Lohr 2007) peuvent être définis similairement.

Les méthodes d'estimation de la variance par rééchantillonnage de bases de sondage multiples peuvent être utilisées avec les logiciels d'analyse de données d'enquête standard qui permettent l'utilisation des poids de rééchantillonnage. Au besoin, chaque colonne de poids de

3. Poststratification sur les chiffres de population de contrôle

Tous les estimateurs décrits à la section 2 modifient les poids de sondage originaux. Par conséquent, certaines propriétés des poids originaux pourraient disparaître. Par exemple, si un échantillon aléatoire stratifié est tiré d'une base de sondage A, les poids modifiés n'auront pas nécessairement la propriété que la somme des poids dans une strate est égale à la taille de population de la strate.

Bankier (1986), dans l'élaboration originale des méthodes d'estimation à base de sondage unique, a proposé de caler par ratisage les poids de sondage, $w_{i,S}^A$ et $w_{i,S}^B$, sur les totaux de strate de manière que les poids ajustés $w_{i,S}^{A,adj}$ et $w_{i,S}^{B,adj}$ satisfassent

$$\sum_{i \in S^{A_h}} (w_{i,S}^{A,adj} + w_{i,S}^{B,adj}) = N^{A_h},$$

où S^{A_h} représente les unités échantillonnées de n'importe quelle base de sondage dans la strate h de la base de sondage A, et N^{A_h} est la taille de la population de cette strate. Bankier (1986) et Skinner (1991) ont utilisé l'estimation par la méthode itérative du quotient (raking ratio) pour caler les estimateurs à base de sondage unique sur les tailles de population des bases de sondage N^A et N^B . Kott, Armrhein et Hicks (1998) ont proposé d'utiliser la méthode de calage par les moindres carrés de Deville et Särndal (1992) pour caler les totaux de population tels que les tailles de strates.

Pour l'estimateur PMV, Lohr et Rao (2000) ont recommandé de combiner d'abord les échantillons, puis d'utiliser les méthodes de calage pour l'ajustement sur le total de population, ainsi que sur les totaux de population des bases de sondage distinctes. En présence de non-réponse et si l'on utilise un estimateur à poids fixes, Brick, Cervantes, Lee et Norman (2011) ont conclu qu'il est préférable de post-stratifier d'abord les échantillons individuels, puis de combiner les échantillons. Dans certaines situations, il est plus efficace de faire la poststratification avant ainsi qu'après avoir combiné les échantillons ; dans d'autres, la poststratification peut accroître le biais (voir section 6). Les décisions concernant la poststratification doivent être fondées sur l'erreur quadratique moyenne, qui comprend les effets des erreurs non dues à l'échantillonnage plutôt que simplement la variance d'échantillonnage.

4. Analyse des données d'enquête à bases de sondage multiples à l'aide de logiciels pour données d'enquête

4.1 Estimation ponctuelle avec des logiciels pour données d'enquête

Seuls les ajustements de poids maintenant la cohérence interne peuvent être utilisés dans les logiciels d'analyse de

données d'enquête s'il existe plusieurs variables d'intérêt. Chacune des méthodes maintenant la cohérence interne présentée à la section 2.1 produisent un vecteur de poids ajustés pour chaque échantillon. Ces vecteurs peuvent ensuite être concaténés pour former un vecteur de poids $\mathbf{w} = [w_1^A, \dots, w_1^B, \dots, w_{i_0}^A, \dots, w_{i_0}^B]$. Soit \mathbf{y} le vecteur correspondant d'observations, formé par la concaténation des observations provenant des échantillons $S(A_1)$ à $S(A_0)$. Alors, $\mathbf{Y} = \mathbf{w}^T \mathbf{y}$. Du point de vue de l'utilisateur, une fois que les poids modifiés sont construits, la procédure suivie pour trouver les estimations ponctuelles des totaux et des moyennes de population est la même que dans le cas d'une enquête à base de sondage unique.

Les poids modifiés provenant d'une procédure avec maintien de la cohérence interne peuvent être utilisés pour estimer toute quantité de population. Soit $F(\mathbf{y})$ la fonction de répartition pour la population, avec

$$F(\mathbf{y}) = \frac{\sum_{i=1}^N I(\mathbf{y}_i \leq \mathbf{y})}{N},$$

où $I(\mathbf{y}_i \leq \mathbf{y}) = 1$ si $\mathbf{y}_i \leq \mathbf{y}$ et 0 autrement. Dans une enquête à base de sondage unique, $F(\mathbf{y})$ est estimée par la fonction de répartition empirique

$$\hat{F}(\mathbf{y}) = \sum_{i \in S} w_i I(\mathbf{y}_i \leq \mathbf{y}) / \sum_{i \in S} w_i.$$

Les poids modifiés peuvent être utilisés pour estimer $F(\mathbf{y})$ dans une enquête à bases de sondage multiples :

$$\hat{F}(\mathbf{y}) = \frac{\sum_{i=1}^0 \sum_{j \in S(A_j)} w_{ij}^A I(\mathbf{y}_i \leq \mathbf{y})}{\sum_{i=1}^0 \sum_{j \in S(A_j)} w_{ij}^A}.$$

Le dénominateur est approximativement sans biais pour N , et le numérateur est approximativement sans biais pour $\sum_{i=1}^N I(\mathbf{y}_i \leq \mathbf{y})$. Toute fonctionnelle de la fonction de répartition peut alors être estimée en utilisant $\hat{F}(\mathbf{y})$: la moyenne, $\int \mathbf{y} d\hat{F}(\mathbf{y})$, la médiane m satisfaisant $F(m) \approx 1/2$, ou toute autre quantité.

Puisque les estimateurs avec poids de sondage modifiés sont approximativement sans biais pour les moyennes et les totaux de population, ils sont aussi approximativement sans biais pour les fonctions lisses des moyennes de population, tels que les ratios et les coefficients de régression. Toute quantité de population qui pourrait être estimée en utilisant les poids provenant d'une enquête à base de sondage unique peut être estimée de manière analogue en utilisant le vecteur de poids ajustés pour l'enquête à bases de sondage multiples.

4.2 Estimation de la variance avec les logiciels pour données d'enquête

Les plans de sondage doivent être connus pour calculer les erreurs-types. L'estimation de la variance est simple

Quand N_{ab} est inconnu, les modifications de la pondération pour les estimateurs PVE sont

$$m_{i,PVE}^a = \begin{cases} \frac{d_{ai}^a}{N_a} \{N_a - N_{PMV}^{ab}(\hat{\theta}^p)\} & \text{si } i \in a \\ \hat{\theta}^p \frac{d_{abi}^a}{N_{PMV}^{ab}} \frac{w_i^a}{d_{abi}^a} \frac{d_{ai}^a}{N_a} \frac{d_{bi}^a}{N_b} \frac{d_{ab}^a}{N_{PMV}^{ab}} (\hat{\theta}^p) & \text{si } i \in ab, \end{cases}$$

$$m_{i,PVE}^b = \begin{cases} \frac{d_{bi}^b}{N_b} \{N_b - N_{PMV}^{ab}(\hat{\theta}^p)\} & \text{si } i \in b \\ (1 - \hat{\theta}^p) \frac{d_{abi}^b}{N_{PMV}^{ab}} \frac{w_i^b}{d_{abi}^b} \frac{d_{bi}^b}{N_b} \frac{d_{ai}^b}{N_a} \frac{d_{ab}^b}{N_{PMV}^{ab}} (\hat{\theta}^p) & \text{si } i \in ab. \end{cases}$$

La contrainte (5) modifie les poids dans le domaine de $S(A)$ soit forcé d'être égal à l'estimateur de X_{ab} pour chevauchement, de sorte que l'estimateur de X_{ab} pour $S(B)$. Toutefois, cette contrainte produit un ensemble différent de poids pour chaque variable de réponse. L'estimateur PVE n'a donc pas de cohérence interne. Rao et Wu (2010) ont présenté une autre version de multiplicité dans laquelle les ajustements des poids ne dépendent pas de y ; en l'absence d'information auxiliaire, cet estimateur est le même que $Y(1/2)$ dans (3).

2.2 Ajustements des poids avec trois bases de sondage ou plus

Dans le cas général, supposons que nous avons \bar{Q} bases de sondage, désignées par $A_1, \dots, A_{\bar{Q}}$. Soit $S(A_q)$ l'échantillon probabiliste tiré de la base de sondage A_q , pour $q = 1, \dots, \bar{Q}$. L'unité i dans l'échantillon $S(A_q)$ a la probabilité d'inclusion π_{i,A_q} et le poids w_{i,A_q} . Il existe un total de D domaines distincts.

Un estimateur pour bases de sondages multiples qui généralise (1) est de la forme

$$Y = \sum_{q=1}^{\bar{Q}} \sum_{i \in S(A_q)} m_{i,A_q}^q w_{i,A_q}^q y_i,$$

où m_{i,A_q}^q est l'ajustement de la pondération pour l'observation i dans $S(A_q)$. Un estimateur à poids fixes établit les ajustements des poids m_{i,A_q}^q pour chaque base de sondage et domaine sous les contraintes que $m_{i,A_q}^q \geq 0$ (m_{i,A_q}^q) est supposé égal à 0 si le domaine d ne fait pas partie de la base de sondage A_q) et $\sum_{q=1}^{\bar{Q}} m_{i,A_q}^q = 1$ pour $d = 1, \dots, D$. Alors, $m_{i,A_q}^q = m_{i,A_q}^q$ quand l'observation i provenant de $S(A_q)$ se trouve dans le domaine d . Un choix simple, qui généralise l'estimateur à deux bases de sondage avec poids fixes $Y(1/2)$ dans (3), se résume à prendre $m_{i,A_q}^q = [1/\text{nombre de bases de sondage qui contiennent le domaine } d]$; il s'agit de l'estimateur fondé sur la multiplicité de Mecatti (2007).

D'autres choix consistent à poser que $m_{i,A_q}^q = 1$ dans exactement une base de sondage et 0 pour les autres bases de sondage, ce qui aboutit aux estimateurs avec sélection.

Nombre de propriétés du cas à deux bases de sondage s'étendent à celui de trois bases de sondage ou plus ; les versions pour bases de sondage multiples des estimateurs de la section 2.1 ont été étudiées par Hartley (1974), Lohr et Rao (2006), et Mecatti (2007). Comment les estimateurs pour bases de sondage multiples satisfont-ils aux critères établis au début de la présente section ? Tous les estimateurs – à poids fixes, optimaux, PMV, PVE et à base de sondage unique – sont approximativement sans biais pour les totaux de population quand des échantillons suffisamment grands sont tirés des bases de sondage. Les estimateurs à poids fixes, PMV et à base de sondage unique ont la propriété de cohérence interne ; par contre, les estimateurs optimaux de type Hartley et de type Fuller-Burneister qui sont donnés dans Lohr et Rao (2006) et une extension aux bases de sondage multiples de l'estimateur PVE de Rao et Wu (2010) ne l'ont pas. Alors que les estimateurs optimaux sont asymptotiquement efficaces, ils sont souvent instables dans les échantillons de taille petite ou moyenne avec trois bases de sondage ou plus, parce que les modifications optimales estimées des poids sont des fonctions de grandes matrices de covariances estimées. Les estimateurs optimaux et les estimateurs PVE conviennent mal pour l'utilisation avec des logiciels d'analyse de données d'enquête standard, parce qu'ils requièrent un ensemble différent de poids pour chaque variable de réponse. Nous recommandons que l'un des estimateurs avec cohérence interne – poids fixes, PMV ou base de sondage unique – soit utilisé en pratique. Lohr et Rao (2006) ont conclu que l'estimateur PMV possède une petite erreur quadratique moyenne dans de nombreuses conditions d'enquête et est donc un bon choix pour une enquête qui n'est réalisée qu'une seule fois. Dans le cas des enquêtes répétées, toutefois, on pourrait privilégier la simplicité et la transparence d'un estimateur à poids fixes. Les ajustements avec poids fixes peuvent faciliter les comparaisons d'une année à l'autre dans le cas d'une enquête annuelle où les proportions de domaine sont relativement constantes au cours du temps. Ces estimateurs se prêtent également mieux aux ajustements de la pondération pour tenir compte de la non-réponse et de la classification incorrecte dans les domaines (voir les sections 5.1 et 6.1). Si l'on peut choisir des ajustements des poids fixes qui sont proches des ajustements de poids optimaux pour les réponses importantes, peut-être en utilisant des effets de plan estimés d'après des enquêtes antérieures, l'estimateur à poids fixes aura une erreur quadratique moyenne proche de celle des estimateurs optimaux et PMV.

Estimateurs du pseudo-maximum de vraisemblance (PMV). Afin d'obtenir la cohérence interne, Skinner et Rao (1996) ont proposé un estimateur du pseudo-maximum de vraisemblance (PMV) qui utilise les mêmes poids pour toutes les variables. Quand N_{ab} est inconnu, il est estimé par $\hat{N}_{PMV}^{ab}(\theta)$, qui est la plus petite des racines de l'équation quadratique

$$\left[\frac{N_B}{\theta} + \frac{1 - \theta}{N_A} \right] x^2 - \left[1 + \theta \frac{N_B}{\hat{N}_A^{ab}} + (1 - \theta) \frac{N_B}{\hat{N}_B^{ab}} \right] x + \theta \hat{N}_A^{ab} + (1 - \theta) \hat{N}_B^{ab} = 0.$$

Skinner et Rao (1996) ont proposé d'utiliser pour θ la valeur $\hat{\theta}_{PMV}$ qui minimise la variance asymptotique de $\hat{N}_{PMV}^{ab}(\theta)$:

$$\hat{\theta}_p = \frac{N_A N_B V(\hat{N}_B^{ab}) + N_B N_A V(\hat{N}_A^{ab})}{N_A N_B V(\hat{N}_B^{ab})} \quad (4)$$

En substituant un estimateur $\hat{\theta}_p$ à θ_p , les ajustements des poids sont :

$$m_{i,p}^A = \begin{cases} \frac{\hat{N}_A^a - \hat{N}_{PMV}^{ab}(\hat{\theta}_p)}{\hat{N}_A^a (\hat{\theta}_p)} & \text{si } i \in a \\ \frac{\hat{\theta}_p \hat{N}_A^{ab} (\hat{\theta}_p)}{\hat{N}_B^b - \hat{N}_{PMV}^{ab}(\hat{\theta}_p)} & \text{si } i \in b \\ \frac{\hat{\theta}_p \hat{N}_A^{ab} + (1 - \hat{\theta}_p) \hat{N}_B^{ab}}{\hat{N}_{PMV}^{ab}(\hat{\theta}_p)} (1 - \hat{\theta}_p) & \text{si } i \in ab. \end{cases}$$

Si la valeur de θ_p ne peut pas être estimée, par exemple si les deux bases de sondage coïncident ou que le plan de la figure 1 est utilisé, on peut se servir d'un effet de plan moyen pour les deux enquêtes dans l'ajustement, comme l'ont décrit Lohr et Rao (2006). L'estimateur PMV est intérieurment cohérent ; même s'il n'est pas garanti qu'il donne l'erreur quadratique moyenne la plus petite, il est d'une très grande efficacité dans de nombreuses situations d'enquête.

Estimateurs à base de sondage unique. Bankier (1986) ainsi que Kalton et Anderson (1986) ont proposé des estimateurs de la forme (1) qui traitent toutes les observations comme si elles avaient été échantillonnées à partir d'une seule base de sondage, en fondant les poids ajustés dans le domaine d'intersection sur les probabilités d'inclusion pour chaque base de sondage. Les ajustements des poids pour l'estimateur à base de sondage unique de Kalton et Anderson (1986) sont :

$$m_{i,S}^A = \begin{cases} 1 & \text{si } i \in a \\ w_B^i / (w_A^i + w_B^i) & \text{si } i \in ab, \\ 1 & \text{si } i \in b \end{cases} \quad m_{i,S}^B = \begin{cases} w_A^i / (w_A^i + w_B^i) & \text{si } i \in ab, \\ 1 & \text{si } i \in b \end{cases}$$

et

$$\sum_{i \in S(A), i \in a} p_{ai}^A = 1, \quad \sum_{i \in S(A), i \in ab} p_{abi}^A = 1, \quad \sum_{i \in S(B), i \in b} p_{bi}^B = 1, \quad \sum_{i \in S(B), i \in ab} p_{abi}^B = 1,$$

(5)

où θ_p est donné par (4). Un estimateur $\hat{\theta}_p$ est substitué à θ_p si ce dernier est inconnu. Alors $\hat{\theta}(\mathbf{p}_a, \mathbf{p}_{ab}, \mathbf{p}_b)$ est maximisé sous les contraintes

$$\hat{\theta}(\mathbf{p}_a, \mathbf{p}_{ab}, \mathbf{p}_b) = \left[\sum_a \frac{N}{N_a + n_B} w_a^i \log(D^a) \right] + \sum_{i \in S(A), i \in ab} \frac{\hat{N}_B^{ab}}{N} w_B^i \log(p_{bi}^B) + \sum_{i \in S(B), i \in b} \frac{\hat{N}_B^b}{N} w_B^i \log(p_{bi}^B) + \sum_{i \in S(A), i \in a} \frac{\hat{N}_A^a}{N} w_A^i \log(p_{ai}^A) + \sum_{i \in S(B), i \in ab} \frac{\hat{N}_B^{ab}}{N} w_B^i \log(p_{abi}^B) \left[(1 - \theta_p) \frac{\hat{N}_B^{ab}}{N} \log(D^B) \right]$$

$\hat{\theta}(\mathbf{p}_a, \mathbf{p}_{ab}, \mathbf{p}_b)$ est définie par vraisemblance empirique (PVE) est définie par sondage. En utilisant $\theta = \theta_p$, la fonction de pseudo-semblance empirique pour les enquêtes à deux bases de Rao et Wu (2010) ont proposé des estimateurs de la vraie

Estimateurs de la pseudo-vraisemblance empirique (PVE). Skinner 1991). (raking) vers les totaux de population de la base de sondage PMV. Leur performance peut être améliorée par ratisage efficaces que les estimateurs optimaux ou les estimateurs base de sondage unique pourraient ne pas être aussi. Toutefois, pour les enquêtes complexes, les estimateurs à que les estimateurs ont la propriété de cohérence interne. sont les mêmes pour toutes les variables de réponse, de sorte Les modifications des poids pour base de sondage unique

dérives, l'estimateur à base de sondage unique se réduit à (3). Les modifications des poids pour base de sondage unique sont les mêmes pour toutes les variables de réponse, de sorte que les estimateurs ont la propriété de cohérence interne. Toutefois, pour les enquêtes complexes, les estimateurs à base de sondage unique pourraient ne pas être aussi efficaces que les estimateurs optimaux ou les estimateurs PMV. Leur performance peut être améliorée par ratisage (raking) vers les totaux de population de la base de sondage (Skinner 1991).

Estimateurs de la pseudo-vraisemblance empirique (PVE). Skinner 1991). (raking) vers les totaux de population de la base de sondage PMV. Leur performance peut être améliorée par ratisage efficaces que les estimateurs optimaux ou les estimateurs base de sondage unique pourraient ne pas être aussi. Toutefois, pour les enquêtes complexes, les estimateurs à que les estimateurs ont la propriété de cohérence interne. sont les mêmes pour toutes les variables de réponse, de sorte Les modifications des poids pour base de sondage unique

Alors, $E[\sum_{i \in S(A)} w_i^A y_i] \approx Y^a + Y^{ab}$ et $E[\sum_{i \in S(B)} w_i^B y_i] \approx Y^b + Y^{ab}$. Par conséquent, un estimateur qui combine les observations provenant des deux enquêtes avec les poids originaux, $\sum_{i \in S(A)} w_i^A y_i + \sum_{i \in S(B)} w_i^B y_i$, est biaisé pour le total de population Y . Si les moyennes de domaines diffèrent, l'estimateur correspondant de la moyenne de population peut également être biaisé.

Les divers estimateurs du total de population Y qui ont été proposés dans la littérature modifient la pondération de façon que les estimateurs soient approximativement sans biais. Les poids modifiés, montrés plus bas pour les divers estimateurs, sont de la forme $\tilde{w}_i^A = m_i^A w_i^A$ et $\tilde{w}_i^B = m_i^B w_i^B$. Le total de population est alors estimé par

$$\hat{Y} = \sum_{i \in S(A)} \tilde{w}_i^A y_i + \sum_{i \in S(B)} \tilde{w}_i^B y_i \quad (1)$$

et la moyenne de population \bar{Y} est estimée par $\hat{\bar{Y}} = \hat{Y} / N$ où

$$N = \sum_{i \in S(A)} \tilde{w}_i^A + \sum_{i \in S(B)} \tilde{w}_i^B.$$

Les estimateurs seront alors approximativement sans biais si $m_i^A \approx 1$ pour $i \in a$, $m_i^B \approx 1$ pour $i \in b$, et $m_i^A + m_i^B \approx 1$ pour $i \in ab$. Tous les estimateurs passés en revue à la présente section satisfont les critères nécessaires pour l'absence approximative de biais en l'absence d'erreurs non dues à l'échantillonnage (voir Lohr 2009).

Ajustements des poids fixes. La modification la plus simple des poids pour préserver l'absence approximative de biais, décrite par Hartley (1962), prend la forme

$$m_{i,\theta}^A = \begin{cases} 1 & \text{si } i \in a \\ \theta & \text{si } i \in ab, \\ 1 - \theta & \text{si } i \in b \end{cases} \quad (2)$$

où $\theta \in [0, 1]$. En utilisant les poids modifiés $\tilde{w}_i^A = m_{i,\theta}^A w_i^A$ et $\tilde{w}_i^B = m_{i,\theta}^B w_i^B$ dans (1), l'estimateur résultant $\hat{Y}(\theta)$ peut aussi être exprimé en se servant des totaux de domaines estimés $\hat{Y}_a^A = \sum_{i \in S(A), i \in a} w_i^A y_i$, $\hat{Y}_b^A = \sum_{i \in S(A), i \in b} w_i^A y_i$, $\hat{Y}_{ab}^A = \sum_{i \in S(A), i \in ab} w_i^A y_i$, et $\hat{Y}_{ab}^B = \sum_{i \in S(B), i \in ab} w_i^B y_i$. L'estimateur

$$\hat{Y}(\theta) = \sum_{i \in S(A)} m_{i,\theta}^A w_i^A y_i + \sum_{i \in S(B)} m_{i,\theta}^B w_i^B y_i$$

$$= \hat{Y}_a^A + \theta \hat{Y}_{ab}^A + (1 - \theta) \hat{Y}_{ab}^B + \hat{Y}_b^B \quad (3)$$

estime donc le total de domaine Y_{ab} par une moyenne pondérée de l'estimateur pour la base de sondage A, \hat{Y}_{ab}^A , et de l'estimateur pour la base de sondage B, \hat{Y}_{ab}^B .

Pour une valeur fixée de θ , l'estimateur $\hat{Y}(\theta)$ assure la cohérence interne, puisque l'on utilise le même ensemble de poids ajustés pour toutes les variables. Cet estimateur est simple à utiliser et à implémenter. Son efficacité dépend de la valeur choisie pour θ . Brick et coll. (2006) ont utilisé $\theta = 1/2$ dans leur étude d'une enquête à deux bases de sondage dans laquelle la base de sondage A est une liste de

sondages avec sélection.

Estimateurs optimaux. Hartley (1962, 1974) a proposé de choisir θ dans (3) de manière que la variance de $\hat{Y}(\theta)$ soit minimisée. La valeur de θ donnant le résultat optimal est

$$\theta_H = \frac{V(\hat{Y}_B^{ab}) + \text{Cov}(\hat{Y}_B^b, \hat{Y}_B^{ab}) - \text{Cov}(\hat{Y}_A^a, \hat{Y}_A^{ab})}{V(\hat{Y}_A^a) + V(\hat{Y}_B^{ab})}.$$

Puisque les variances et les covariances sont généralement inconnues, elles doivent être estimées d'après les données, ce qui donne

$$\hat{\theta}_H = \frac{V(\hat{Y}_B^{ab}) + \text{Cov}(\hat{Y}_B^b, \hat{Y}_B^{ab}) - \text{Cov}(\hat{Y}_A^a, \hat{Y}_A^{ab})}{V(\hat{Y}_A^a) + V(\hat{Y}_B^{ab})}.$$

Skinner et Rao (1996) ont montré que l'estimateur de Hartley peut être calculé en utilisant des poids ajustés. Les modifications des poids pour l'estimateur de Hartley $\hat{Y}(\theta^H)$ sont données par (2), en substituant θ^H à θ . Puisque θ^H est convergent pour θ^H , l'estimateur de Hartley est asymptotiquement optimal parmi les estimateurs de la forme $\hat{Y}_a^A + \hat{Y}_b^B + \theta \hat{Y}_{ab}^A + (1 - \theta) \hat{Y}_{ab}^B$. Cependant, les poids modifiés $\tilde{w}_{i,H}^A$ et $\tilde{w}_{i,H}^B$ sont des fonctions des variances et des covariances des totaux de domaines estimés. Cela a deux conséquences : 1) les poids modifiés sont des variables aléatoires et leur variabilité doit être prise en compte dans les erreurs-types des estimateurs et 2) les modifications optimales des poids varieront pour différentes variables de réponse, ce qui donnera lieu à une incohérence interne.

Fuller et Burmeister (1972) ont proposé de modifier l'estimateur de Hartley en se servant d'information additionnelle au sujet de N^{ab} , ce qui donne

$$\hat{Y}_{FB}^{ab}(\beta) = \hat{Y}_a^A + \hat{Y}_b^B + \beta_1 \hat{Y}_{ab}^A + (1 - \beta_1) \hat{Y}_{ab}^B + \beta_2 (\hat{N}_a^A - \hat{N}_b^B).$$

Comme dans le cas de l'estimateur de Hartley, les valeurs optimales β_1^{opt} et β_2^{opt} sont choisies de manière à minimiser la variance de $\hat{Y}_{FB}^{ab}(\beta)$, et sont donc des fonctions des covariances des totaux de domaines. En substituant les estimateurs convergents $\hat{\beta}_1^{\text{opt}}$ et $\hat{\beta}_2^{\text{opt}}$, on obtient les ajustements des poids pour w_i^A et w_i^B . Lohr et Rao (2000) ont montré que l'estimateur de Fuller-Burmeister \hat{Y}_{FB} est celui qui, de tous les estimateurs considérés, possède la plus petite variance asymptotique. Comme dans le cas de l'estimateur de Hartley, cependant, les poids modifiés sont des variables aléatoires qui diffèrent pour des réponses différentes et, dans les plans d'échantillonnage complexes, l'estimateur de Fuller-Burmeister manque aussi de cohérence interne.

d'après l'information combinée. À la section 5, nous discutons des effets de la non-réponse et des effets de mode dans les enquêtes à bases de sondage multiples. En plus des problèmes de non-réponse, de sous-dénombrement et d'erreurs de mesure qui se posent dans le cas des enquêtes à base de sondage unique, les enquêtes à bases de sondage multiples peuvent poser le problème de la classification incorrecte dans les domaines. La modification des poids pour les estimateurs de la section 2 dépend de l'appartenance des observations aux domaines. S'il est probable que certaines observations appartenant au domaine a soient enregistrées par erreur comme appartenant au domaine ab , les estimateurs pourraient présenter un biais important. Nous étudions les effets de l'erreur de classification dans les domaines à la section 6 et proposons une nouvelle méthode pour corriger le biais dû à l'erreur de classification quand les probabilités d'erreur de classification sont connues. Enfin, à la section 7, nous examinons les questions concernant le plan de sondage et à la section 8, nous discutons des possibilités et des défis associés aux enquêtes à bases de sondage multiples.

2. Estimateurs dans les enquêtes à bases de sondage multiples chevauchantes

À la présente section, nous passons en revue les estimateurs du total de population Y pour les enquêtes à bases de sondage multiples chevauchantes, ainsi que les modifications de la pondération induite par ces estimateurs. Pour simplifier la notation, nous nous concentrons sur une enquête à deux bases de sondage à la section 2.2. Dans une enquête à deux bases de sondage, nous pouvons écrire

$$Y = Y_a + Y_{ab} + Y_b,$$

où Y_a est le total des unités de la population dans le domaine a , Y_{ab} est le total des unités de la population dans le domaine ab , et Y_b est le total des unités de la population dans le domaine b . Un cas particulier est l'estimation de la taille de population $N = N_a + N_{ab} + N_b$, dont il est question dans Haines et Pollock (1998). Nous discutons de l'estimation d'autres quantités de population que les totaux et les moyennes, et utilisons les données provenant d'enquête à bases de sondage multiples dans d'autres analyses, à la section 4. Nous commençons par énoncer certaines propriétés souhaitables des estimateurs pour enquêtes à bases de sondage multiples.

1. Un estimateur devrait être approximativement sans biais pour la quantité de population finie correspondante.
2. Les estimateurs devraient avoir la propriété de cohérence interne : autrement dit, si X_1 estime le nombre de femmes ingénieures dans la population, X_2 estime le nombre d'hommes ingénieurs dans la population et X_3 estime le nombre total d'ingénieurs dans la population, nous devrions avoir $X_1 + X_2 = X_3$. La cohérence interne préserve les relations multivariées dans les données. Pratiquement parlant, la cohérence interne requiert qu'un seul ensemble de poids soit utilisé pour toutes les estimations.
3. Un estimateur devrait être efficace, c'est-à-dire avoir une faible erreur quadratique moyenne.
4. Un estimateur devrait avoir une forme permettant de le calculer au moyen des logiciels standards tels que SUDAAN ou SAS PROC SURVEYMEANS. Cela permet aux analystes de travailler avec les données sans devoir écrire ou tester de nouveaux logiciels. Du point de vue pratique, un seul fichier de données est créé pour l'enquête à bases de sondage multiples. Le fichier comprend une colonne de poids qu'il faut utiliser pour calculer les estimations ponctuelles, et il contient soit des variables décrivant les plans de sondage pour l'estimation de la variance fondée sur une formule, ou des colonnes de poids de rééchantillonnage pour l'estimation de la variance par les méthodes de rééchantillonnage.
5. Un estimateur devrait, dans la mesure du possible, être robuste aux erreurs non dues à l'échantillonnage qui pourraient survenir dans les enquêtes à bases de sondage multiples.

2.1 Estimateurs et ajustement de la pondération pour les enquêtes à deux bases de sondage

Considérons l'enquête à deux bases de sondage chevauchantes illustrée à la figure 2, où le domaine ab n'est pas vide. Un échantillon probabiliste $S(A)$ de taille n_a est tiré des N_a unités dans la base de sondage A , et un échantillon probabiliste indépendant $S(B)$ de taille n_b est tiré des N_b unités de la base de sondage B . L'unité i dans l'échantillon $S(A)$ possède la probabilité d'inclusion π_i^A et le poids w_i^A , et l'unité j dans l'échantillon $S(B)$ possède la probabilité d'inclusion π_j^B et le poids w_j^B . Les poids peuvent être l'inverse des probabilités d'inclusion, ou ils peuvent être poststratifiés afin qu'ils concordent avec les chiffres de population ; nous supposons que les estimateurs des totaux de population sont approximativement sans biais.

que les bases de sondage chevauchantes B et C sont toutes deux incomplètes, mais moins coûteuses à échantillonner. Ce plan de sondage a été utilisé aux États-Unis pour les enquêtes du *Scientists and Engineers Statistical Data System* (SESTAT ; *National Science Foundation* 2003). Le même plan de sondage pourrait être utilisé quand A est la base de sondage pour une enquête de population générale, B est une enquête auprès des possesseurs d'un téléphone fixe et C, l'enquête auprès des possesseurs d'un téléphone mobile.

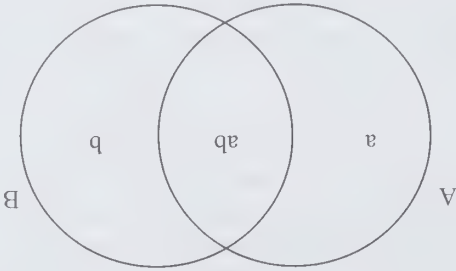


Figure 2 Les bases de sondage A et B se chevauchent, créant les trois domaines a, b et ab

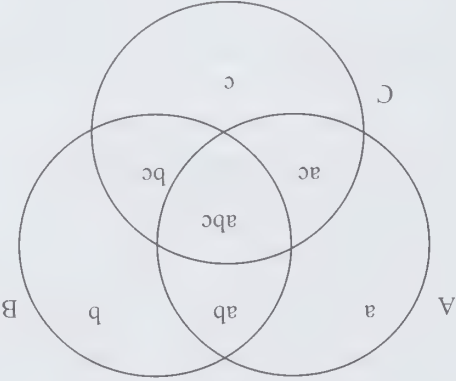


Figure 3 Les bases de sondage A, B et C sont toutes les trois incomplètes et se chevauchent

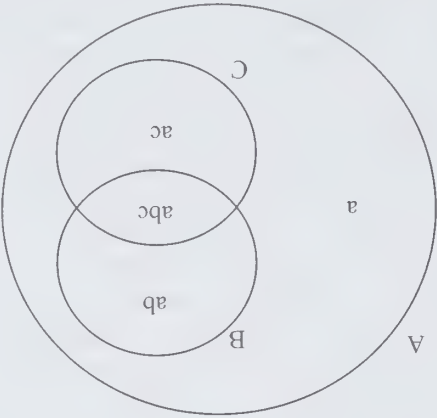


Figure 4 La base de sondage A contient l'entière de la population ; les bases de sondage B et C se chevauchent et sont toutes les deux contenues dans la base de sondage A

L'utilisation de plans à bases de sondage multiples dans les enquêtes-ménages offre de nombreuses possibilités, dont :

1. L'utilisation de bases de sondage multiples fondées sur des dossiers administratifs ;
2. L'échantillonnage à modes multiples (par exemple, le tirage d'échantillons indépendants à partir d'une liste de numéros de téléphone mobile et à partir d'une liste de numéros de téléphone fixe) ;
3. L'utilisation future d'Internet pour la collecte des données. Quoique Internet pose de nombreux problèmes de couverture et de spécification des domaines, son utilisation mérite d'être envisagée, en raison des économies possibles et de la facilité de recueillir et de traiter les données ;
4. L'amélioration de l'estimation sur petits domaines. Une enquête nationale peut être complétée par des enquêtes locales, plus petites, pour obtenir des estimations de plus grande précision dans les petits domaines ;
5. L'amélioration de l'estimation pour des populations rares. Une enquête de population générale peut être complétée par une enquête auprès d'un échantillon tiré d'une base de sondage dans laquelle la concentration des membres d'une population rare est élevée ;
6. L'emploi de plans de sondage modulaires. Une approche à bases de sondage multiples peut donner plus de souplesse pour la conception des enquêtes permanentes. À mesure que l'échantillonnage à partir de bases de sondage particulières devient moins cher, la répartition relative de la taille de l'échantillon entre les diverses bases de sondage peut être modifiée. L'approche modulaire donne aussi plus de souplesse pour répondre à l'évolution des besoins de données.

La plus grande souplesse des enquêtes à bases de sondages multiples a toutefois pour prix l'accroissement de la complexité. L'information provenant des diverses enquêtes doit être combinée pour estimer les quantités de population et les estimateurs que l'on peut choisir sont nombreux. À la section 2, nous résumons les estimateurs qui ont été élaborés pour les totaux de population et décrivons comment les poids de sondage sont modifiés dans ces estimateurs ; aux sections 3 et 4, nous discutons du calage des poids et décrivons comment utiliser les logiciels d'analyse de données d'enquête avec des données provenant d'enquêtes à bases de sondage multiples. Les erreurs non dues à l'échantillonnage doivent être prises en considération en ce qui concerne leur effet sur les estimations calculées séparément ainsi que leur effet sur les estimations calculées

échantillonnage sera efficace ; l'échantillon tiré de la base de sondage A, quoique plus coûteux, donne une couverture complète de la population.

Dans d'autres situations, toutes les bases de sondage sont incomplètes, comme l'a considéré Hartley (1962) ; par exemple, à la figure 2, la base de sondage A pourrait être une liste de numéros de téléphone fixe et la base de sondage B, une liste de numéros de téléphone mobile. Il existe trois domaines : le domaine *a* contient les unités figurant dans la base de sondage A mais non dans la base de sondage B, le domaine *b* contient les unités figurant dans la base de sondage B mais non dans la base de sondage A, et le domaine *ab* contient les unités figurant dans les deux bases de sondage. Dans le contexte téléphonique, le domaine *a* contient les individus appartenant à un ménage ne possédant qu'un téléphone fixe, le domaine *b* contient les individus possédant qu'un téléphone mobile et le domaine *ab* comprend les individus qui ont à la fois un téléphone mobile et un téléphone fixe. On ne sait pas d'avance si un membre du ménage échantillonné en utilisant l'une des bases de sondage appartient aussi à l'autre (Brick, Dipko, Presser, Tucker et Yuan 2006) ; habituellement, on demande aux enquêtés des renseignements sur leur utilisation d'un téléphone mobile et d'un téléphone fixe pour déterminer le domaine auquel ils appartiennent.

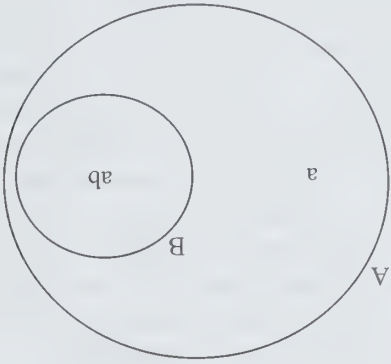


Figure 1 Un plan à deux bases de sondage dans lequel la base de sondage B est un sous-ensemble de la base de sondage A

Plus de deux bases de sondage peuvent être utilisées également, comme l'illustre la figure 3 pour une enquête à trois bases de sondage dans laquelle toutes les bases de sondage sont incomplètes. Dans cette situation, il existe sept domaines. Iachan et Dennis (1993) ont donné un exemple d'enquête à trois bases de sondage utilisée pour échantillonner la population de sans-abri, où la base de sondage A est une liste des refuges et la base de sondage B est une liste des refuges et la base de sondage C correspond à des emplacements de rue. La figure 4 représente une enquête à trois bases de sondage dans laquelle la base de sondage A donne une couverture complète, tandis

L'une des premières enquêtes à bases de sondage multiples (à part les premières méthodes de capture-recapture) a été menée par le Census Bureau en 1949 (Hansen, Hurwitz et Madow 1953). Dans la Sample Survey of Retail Stores, un échantillon probabiliste d'unités primaires d'échantillonnage (UPE) a été sélectionné. Dans chaque UPE, une liste des grandes entreprises de commerce de détail a été établie d'après les dossiers de l'Old Age and Survivors Insurance Bureau. Toutes les entreprises figurant sur la liste ont été incluses dans l'échantillon et un échantillon aréolaire des UPE ne figurant pas sur la liste a été sélectionné. Dans ce cas, un plan à deux bases de sondage avec sélection a été employé dans chaque UPE sélectionnée ; les unités figurant sur la liste ont été retirées de la base aréolaire avant l'échantillonnage. Donc, dans chaque UPE, l'estimateur du total des ventes était égal à la somme des deux estimateurs. Aucune nouvelle méthode statistique n'était nécessaire pour estimer le total des ventes dans cette enquête, puisqu'essentiellement, un échantillon stratifié était tiré dans chaque UPE : les entreprises des UPE reprises sur la liste formaient une strate et celles figurant dans la base aréolaire mais non sur la liste formaient la deuxième strate. Cette méthode d'enquête a permis de réaliser des économies, parce qu'il était relativement peu coûteux d'échantillonner les entreprises sur la liste, tout en obtenant une couverture complète grâce à l'utilisation de la base aréolaire.

De nombreuses enquêtes agricoles se sont également appuyées sur un plan à deux bases de sondage avec sélection (González-Villalobos et Wallace 1996). Dans ce type de plan, les fermes figurant sur la liste sont retirées de la base aréolaire avant que l'échantillonnage ne commence. Des économies importantes peuvent être réalisées, car l'échantillonnage d'après la liste est souvent nettement moins coûteux et que la liste contient les plus grandes entreprises.

Toutefois, dans de nombreux cas, il n'est pas forcément possible ni pratique de supprimer de la base aréolaire avant l'échantillonnage les unités qui figurent sur la liste. Au lieu de cela, dans une enquête à deux bases de sondage chevauchantes, des échantillons probabilistes indépendants sont tirés de la base de sondage A (la base de sondage aréolaire) et de la base de sondage B (la liste), comme l'illustre la figure 1. Les populations rares peuvent souvent être échantillonnées plus efficacement en se servant d'un échantillon à bases de sondage multiples (Kallion et Anderson 1986). Par exemple, dans une étude épidémiologique, la base de sondage A pourrait être utilisée pour une enquête générale sur la santé de la population, tandis que la base de sondage B pourrait être une liste des cliniques qui se spécialisent dans une maladie particulière. L'échantillon tiré de la base de sondage B devrait contenir un pourcentage élevé de personnes atteintes de la maladie d'intérêt de sorte que cet

Autres plans de sondage : échantillonnage avec bases de sondage chevauchantes

Sharon L. Lohr¹

Résumé

Les plans de sondage et les estimateurs des enquêtes à base de sondage unique utilisés à l'heure actuelle par les organismes gouvernementaux américains ont été élaborés en réponse à des problèmes pratiques. Les programmes d'enquêtes-ménages fédéraux doivent faire face aujourd'hui à la diminution des taux de réponse et de la couverture des bases de sondage, à la hausse des coûts de collecte des données et à l'accroissement de la demande de statistiques pour des petits domaines. Les enquêtes à bases de sondage multiples, dans lesquelles des échantillons indépendants sont tirés de bases de sondage distinctes, peuvent être utilisées en vue de relever certains de ces défis. La combinaison d'une liste et d'une base de sondage aréolique ou l'utilisation de deux bases de sondage pour échantillonner les ménages ayant une ligne de téléphone fixe et ceux ayant une ligne de téléphone mobile en sont des exemples. Nous passons en revue les estimateurs ponctuels et les ajustements de la pondération qui peuvent être utilisés pour analyser les données d'enquête à bases de sondage multiples au moyen de logiciels standard et nous résumons la construction des poids de rééchantillonnage pour l'estimation de la variance. Étant donné leur complexité croissante, les enquêtes à bases de sondage multiples obligent à résoudre des difficultés qui ne se posent pas dans le cas des enquêtes à base de sondage simple. Nous étudions le biais dû à l'erreur de classification dans les enquêtes à bases de sondage multiples et proposons une méthode pour corriger ce biais quand les probabilités d'erreur de classification sont connues. Enfin, nous discutons des travaux de recherche nécessaires en ce qui concerne les erreurs non dues à l'échantillonnage dans les enquêtes à bases de sondage multiples.

Mots clés : Correction du biais ; enquête à deux bases de sondage ; erreur de classification ; effets de mode ; échantillonnage d'événements rares ; poids d'échantillonnage ; estimation sur petits domaines.

1. Utilisation d'enquête à bases de sondage multiples

En théorie classique de l'échantillonnage fondé sur un plan de sondage, un échantillon probabiliste est tiré de la base de sondage (unique) et les probabilités d'inclusion dans le plan de sondage peuvent être utilisées pour faire des inférences au sujet de la population. Soit y_i une mesure sur l'unité i dans la population de N unités, soit S l'ensemble d'unités dans l'échantillon, et soit $\pi_i = P$ (l'unité i est incluse dans l'échantillon). Alors, l'estimateur de Horvitz-Thompson (1952) du total de population $Y = \sum_{i=1}^N y_i$ est $\hat{Y} = \sum_{i \in S} w_i y_i$, où $w_i = 1/\pi_i$ est le poids de sondage. Si la base de sondage englobe toutes les unités faisant partie de la population cible, que toutes les unités échantillonnées répondent et qu'il n'y a pas d'erreur de mesure, l'estimateur de Horvitz-Thompson est sans biais pour Y .

Les difficultés pratiques posées par l'échantillonnage durant les années 1940 et les années 1950 ont mené les méthodologistes à élaborer des sondages et des estimateurs stratifiés à plusieurs degrés tels que l'estimateur de Horvitz-Thompson. Les enquêtes avec interview sur place s'appuyaient sur un échantillonnage avec probabilités inégales pour équilibrer les charges de travail des intervieweurs et réduire les variances. Les taux de réponse étaient élevés dans le cas de nombreuses enquêtes gouvernementales, de sorte que toutes les hypothèses qui sous-tendent l'estimateur de Horvitz-Thompson étaient raisonnables. Aujourd'hui,

nous devons faire face à de nouvelles difficultés dans les enquêtes-ménages. Les taux de non-réponse augmentent, si bien que les estimations par sondage doivent s'appuyer davantage sur des modèles. La diversité ethnique et linguistique d'une population peut donner lieu à un sous-dénombrement et à une erreur de mesure. La diversité technologique croissante signifie qu'il est parfois préférable de recourir à des modes d'échantillonnage différents pour divers types de résidents ; on doit alors être certain que les divers modes d'échantillonnage mesurent les mêmes quantités. Les coûts de la collecte des données ont grimpé fortement, en partie à cause de l'accroissement de la non-réponse ; parallèlement, les demandes de données des organismes gouvernementaux et des chercheurs ont également augmenté considérablement.

Les enquêtes à bases de sondage multiples permettent d'obtenir une meilleure couverture de la population à un coût plus faible. Elles peuvent être utilisées comme un élément structurel d'un plan de sondage modulaire qui s'appuie sur différentes bases de sondage pour essayer de réduire les coûts et accroître la couverture. Elles permettent aussi d'utiliser efficacement les données administratives. Dans le présent article, nous décrivons divers types d'enquêtes à bases de sondage multiples et discutons de certains travaux de recherche achevés et de ceux qui pourraient être nécessaires pour que l'on puisse utiliser ces enquêtes.

Remerciements

L'auteur remercie le Census Bureau, en particulier Patrick Flanagan, d'avoir été l'hôte d'une série d'exposés magistraux sur lesquels le présent article est fondé en partie. La recherche sur les plans d'échantillonnage décrits a été financée par le Conseil de recherches en sciences naturelles et en génie du Canada, la National Science Foundation, le National Center for Health Statistics, les Centers for Disease Control and Prevention, les Los Alamos National Laboratories et les National Institutes of Health.

Bibliographie

- Basu, D. (1969). Role of the sufficiency and likelihood principles in sample survey theory. *Sankhyā*, 31, 441-454.
- Birnbaum, Z.W., et Sirken, M.G. (1965). Design of sample surveys to estimate the prevalence of rare diseases: Three unbiased estimators. *Vital and Health Statistics*, Série 2, No. 11. Washington : Government Printing Office.
- Brin, S., et Page, L. (1998). The anatomy of a large-scale hypertextual web search engine. *Proceedings of the 7th International World Wide Web Conference*. Elsevier, 107-117.
- Casella, G., et Robert, C.P. (1996). Rao-Blackwellization of sampling schemes. *Biometrika*, 83.
- Chao, C.-T., et Thompson, S.K. (2001). Optimal adaptive selection of sampling sites. *Environmetrics*, 12, 517-538.
- Chow, M., et Thompson, S.K. (2003). Estimation avec plans d'échantillonnage par dépistage de liens – Une approche bayésienne. *Techniques d'enquête*, 20, 221-230.
- Felix-Medina, M.H., et Thompson, S.K. (2004). Combining link-tracing sampling and cluster sampling to estimate the size of hidden populations. *Journal of Official Statistics*, 20, 19-38.
- Frank, O. (1971). *Statistical Inference in Graphs*. Stockholm : Försvarets Forskningsanstalt.
- Frank, O. (1977a). Survey sampling in graphs. *Journal of Statistical Planning and Inference*, 1, 235-264.
- Frank, O. (1977b). Estimation of graph totals. *Scandinavian Journal of Statistics*, 4, 81-89.
- Frank, O. (1978a). Estimating the number of connected components in a graph by using a sampled subgraph. *Scandinavian Journal of Statistics*, 5, 177-188.
- Frank, O. (1978b). Sampling and estimation in large social networks. *Social Networks*, 1, 91-101.
- Frank, O. (1979). Estimation of population totals by use of snowball samples. Dans *Perspectives on Social Network Research*, (Eds., P.W. Holland et S. Leinhardt). New York : Academic Press, 319-347.
- Frank, O., et Snijders, T. (1994). Estimating the size of hidden populations using snowball sampling. *Journal of Official Statistics*, 10, 53-67.
- Thompson : Échantillonnage adaptatif par réseau et spatial
- Hasings, W.K. (1970). Monte-Carlo sampling methods using Markov chains and their application. *Biometrika*, 57, 97-109.
- Heckathorn, D.D. (1997). Respondent driven sampling: A new approach to the study of hidden populations. *Social Problems*, 44, 174-199.
- Heckathorn, D.D. (2002). Respondent driven sampling II: Deriving valid population estimates from chain-referral samples of hidden populations. *Social Problems*, 49, 11-34.
- Klov Dahl, A.S. (1989). Urban social networks: Some methodological problems and possibilities. Dans *The Small World* (Ed., T. Szöni). János Bolyai Mathematical Society, Keszthely, Hungary, 2, 1-46.
- Salganik, M.J., et Heckathorn, D.D. (2004). Sampling and estimation in hidden populations using respondent-driven sampling. *Sociological Methodology*, 34, 193-239.
- Sirken, M.G. (1970). Household surveys with multiplicity. *Journal of the American Statistical Association*, 63, 257-266.
- Sirken, M.G. (1972a). Stratified sample surveys with multiplicity. *Journal of the American Statistical Association*, 67, 224-227.
- Sirken, M.G. (1972b). Variance components of multiplicity estimators. *Biometrics*, 28, 869-873.
- Thompson, S.K. (1990). Adaptive cluster sampling. *Journal of the American Statistical Association*, 85, 1050-1059.
- Thompson, S.K. (2002). *Sampling, Second Edition*. New York : John Wiley & Sons, Inc.
- Thompson, S.K. (2006a). Plans de sondage à marche aléatoire ciblée. *Techniques d'enquête*, 32, 11-26.
- Thompson, S.K. (2006b). Adaptive web sampling. *Biometrics*, 62, 1224-1234.
- Thompson, S., et Frank, O. (2000). Estimation fondée sur un modèle et comportant des plans d'échantillonnage à dépistage de liens. *Techniques d'enquête*, 26, 99-112.
- Thompson, S.K., et Seber, G.A.F. (1996). *Adaptive Sampling*. New York : John Wiley & Sons, Inc.
- Zacks, S. (1969). Bayes sequential designs of fixed size samples from finite populations. *Journal of the American Statistical Association*, 64, 1342-1349.

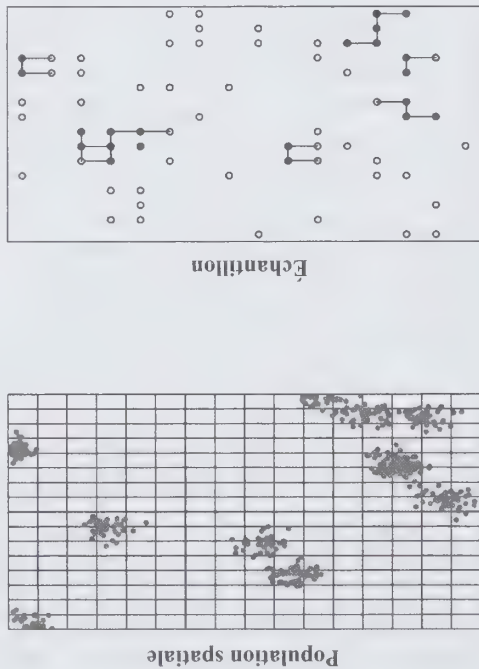
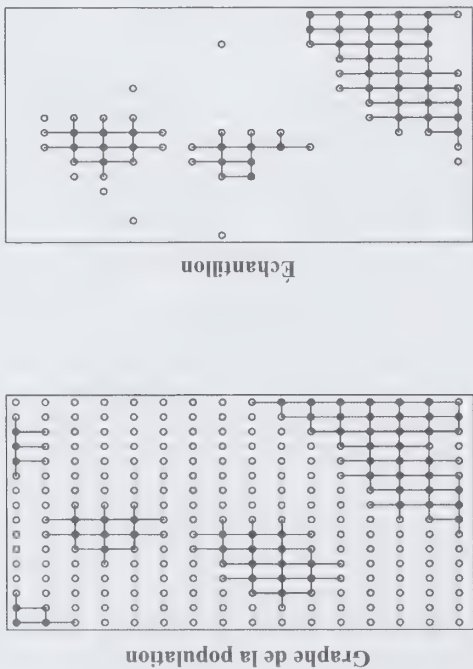


Figure 17 Variantes de plan d'échantillonnage adaptatif « en toile »



Les plans d'échantillonnage adaptatifs élargissent considérablement le champ des possibilités en matière de stratégies d'échantillonnage. Ils semblent être particulièrement utiles pour les populations qui sont difficiles à échantillonner par d'autres moyens. Les plans d'échantillonnage par réseau sont intrinsèquement adaptatifs dans la plupart des cas et peuvent offrir des moyens plus efficaces d'échantillonner les populations présentant une structure en réseau et une structure spatiale. Dans le présent article, l'accent est mis sur les plans produisant une faible erreur quadratique moyenne ou fournissant des moyens pratiques d'atteindre un objectif pourrait être simplement d'obtenir un échantillon offrant un plus haut rendement, c'est-à-dire un échantillon pour lequel la valeur totale de la variable d'intérêt est élevée. Par exemple, les points chauds environnementaux sont ceux où doit avoir lieu l'assainissement, les composantes à haut risque d'une épidémie reliée à un réseau sont celles où le traitement ou l'intervention pourrait avoir l'effet le plus important. Les avantages d'une approche adaptative sont encore plus directs quand l'objectif est d'obtenir un haut rendement d'échantillon.

Les stratégies d'échantillonnage entièrement optimales ne sont, dans la plupart des cas, pas pratiques à mettre en œuvre, à cause de la difficulté des calculs et de la dépendance à l'égard d'un modèle. Une approche plus pratique consiste à apporter des améliorations à des plans de sondage classiques au moyen de procédures adaptatives simples qui

4. Discussion

saisissent en grande partie l'essence du problème, et le choix d'un plan de sondage a souvent nettement plus d'effet que celui d'une méthode d'inférence plutôt qu'une autre. Les analyses en simulation s'appuyant sur des stratégies adaptatives de différents types ont eu tendance à appuyer la notion qu'il est bon d'avoir une forte composante classique sous-jacente. Nombre des stratégies pratiques prennent la forme d'un échantillon classique initial conjugué à un échantillonnage adaptatif pour étendre l'échantillon en partant de ce point à travers des relations en réseau ou spatiales et en fonction des valeurs observées. Les stratégies comportant ce genre d'équilibre entre les composantes classiques et adaptatives ont en général donné de meilleurs résultats dans les simulations que, disons, la sélection d'une unité unique classiquement et l'ajout adaptatif de tout le reste de l'échantillon à partir de là. Dans les simulations, les stratégies les plus efficaces ont tendance à comprendre un échantillonnage initial représentant environ 60 % à 80 % de la taille totale de l'échantillon. La quantité modeste d'échantillonnage adaptatif après cela produit alors d'importants gains d'efficacité. Cette expérience empirique est en harmonie avec les caractéristiques des stratégies adaptatives optimales, dans lesquelles il semble exister des tiraillements entre l'étalement des unités à grande distance ou le remplissage des parties non observées de la région étudiée, ce qui correspond à la composante classique des plans de sondage simplifiés, et placer de nouvelles unités dans les régions les plus prometteuses, ce qui correspond à la composante adaptative dans les plans de sondage simplifiés.

atteignant aussi les composantes. Dans le plan illustré à droite, une unité initiale unique est sélectionnée au hasard et l'échantillonnage adaptatif « en toile » se poursuit jusqu'à l'obtention d'un total de 80 unités. La probabilité de 0,1 de sélectionner l'unité suivante au hasard à n'importe quelle étape empêche que le plan soit calé dans n'importe laquelle des composantes. Sous ce plan, les composantes ou agrégats principales sont couvertes de façon très détaillée, quoique non exhaustive.

3.1 Plans spatiaux avec liens pondérés

Pour sélectionner des échantillons spatiaux, les poids des liens peuvent être définis sous forme d'une fonction de la distance entre les emplacements. Par exemple, pour un échantillon augmenté, la fonction attribuerait des poids plus grands aux emplacements situés à courte distance. Par ailleurs, pour remplir l'espace, les emplacements situés à plus grande distance recevraient des poids plus grands. Un plan d'échantillonnage par réseau dans ces conditions, avec des poids de lien définis uniquement sur la base de distance, ne serait généralement pas adaptatif. Cela est dû au fait que la base de sondage spatiale permettrait, en se servant d'un plan par dépistage de liens, de sélectionner l'échantillon complet d'emplacements avant de se rendre sur le terrain pour faire des observations.

Toutefois, de manière plus générale, les poids des liens peuvent être définis sous forme d'une fonction des poids dans l'échantillon, possédant une valeur observée élevée de la variable d'intérêt, la fonction pourrait attribuer un poids plus élevé aux emplacements proches et un poids plus faible aux emplacements éloignés. Pour l'unité ayant une valeur faible de la variable d'intérêt, la fonction de poids pourrait avoir une forme plus uniforme.

Les plans de sondage à marche aléatoire sont particulièrement faciles à exécuter dans des conditions spatiales avec des poids de lien indépendants de la distance. Cela tient au fait que, à tout point de l'échantillonnage, la sélection de l'emplacement suivant ne dépend que de l'emplacement sélectionné le plus récemment, de sorte que l'on ne doit prendre en considération qu'une seule fonction de poids. Dans le cas de plans plus généraux, tels que l'échantillonnage adaptatif « en toile », l'utilisation de fonctions de poids de lien qui dépendent à la fois de la distance et de la valeur offre une très grande souplesse en ce qui a trait aux stratégies adaptatives possibles.

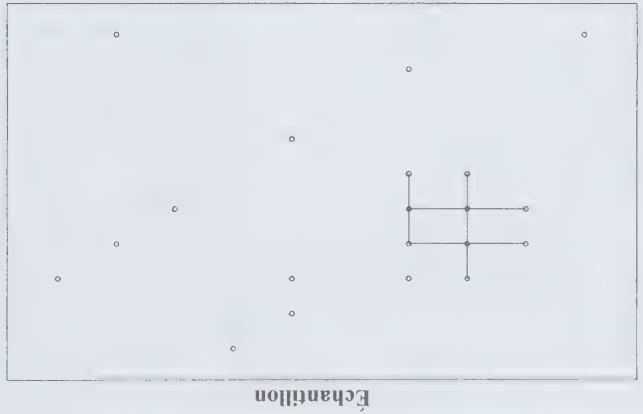


Figure 15 Échantillon adaptatif « en toile » de 20 unités partant de l'échantillon initial de la figure précédente

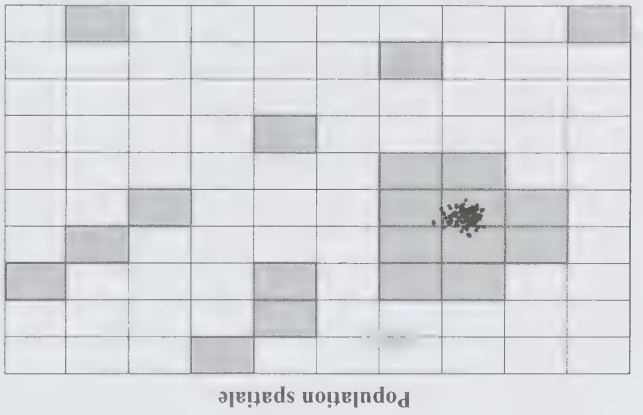


Figure 16 Représentation spatiale de l'échantillon adaptatif en toile

Un aperçu de l'immense souplesse qu'offrent les plans d'échantillonnage adaptatif « en toile » dans le contexte spatial est donné à la figure 17. À la rangée supérieure, une population spatiale est convertie en un graphe, quoique les directions des liens ne soient pas montrées. La rangée inférieure présente des échantillons provenant de deux variantes de l'échantillonnage adaptatif « en toile ». À gauche, 16 unités initiales ont été sélectionnées indépendamment au hasard. À partir de chacune, une procédure d'échantillonnage « en toile » adaptative est exécutée jusqu'à l'obtention d'une taille d'échantillon de cinq unités. Sous ce plan, l'échantillon est réparti à travers la région étudiée, tout en

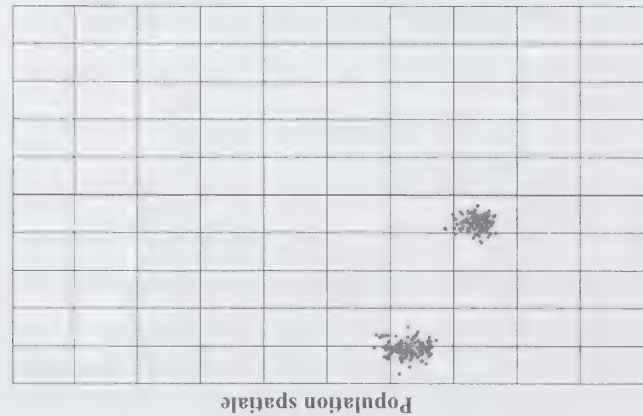


Figure 12 Une population spatialement regroupée

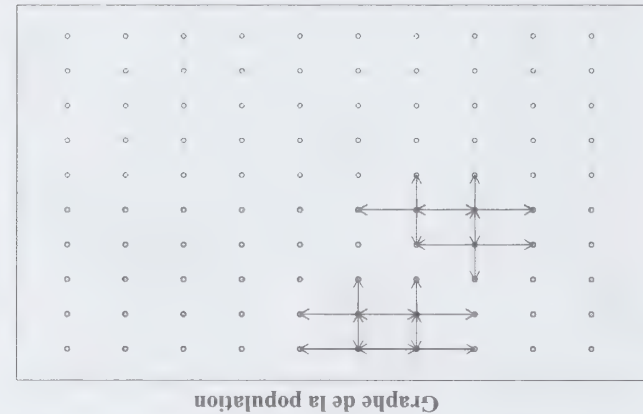


Figure 13 Un réseau représentant les relations de voisinage pertinentes dans la population spatiale



Figure 14 Un échantillon aléatoire initial d'unités spatiales

Dans le cas de l'échantillonnage adaptatif par grappes, la contrainte consistant à continuer d'échantillonner jusqu'à ce que tous les voisins de toutes les unités satisfaisant la condition aient été inclus signifie que la taille globale d'échantillon n'est pas contrôlée d'avance et elle est plutôt astreignante quand certains réseaux sont inhabituellement grands. L'échantillonnage adaptatif « en toile » dans le contexte spatial résout ce problème puisque la taille d'échantillon peut être fixée d'avance. Pour ce qui est de sa redéfinition du réseau, les simples estimateurs sans biais de l'échantillonnage adaptatif par grappes emploient des données provenant uniquement des composantes fortement connectées que recoupe l'échantillon initial. Les améliorations de type Rao-Blackwell fondées sur ces estimateurs peuvent en outre utiliser des données provenant des extensions faiblement connectées de ces composantes. Les unités de liste (*edge units*) bien connues de l'échantillonnage adaptatif par grappes spatial sont un cas particulier de ces extensions faiblement connectées de composantes fortement connectées.

La figure 12 illustre une région à l'étude contenant une population regroupée spatialement comme on peut en observer dans les enquêtes écologiques, épidémiologiques et sociodémographiques. Dans une forme de plan de sondages spatiaux adaptatifs, le voisinage d'une unité est défini comme l'ensemble d'unités directement adjacentes, et les unités voisines sont ajoutées à l'échantillon quand la valeur d'une unité échantillonnée est élevée ou satisfait un autre critère donné. À la figure 13, la population spatiale a été transformée en un graphe orienté. Les unités spatiales cartées sont redessinées comme des nœuds dans un graphe, et chaque fois que le nombre d'objets dans une unité est supérieur à zéro, des flèches représentant les liens du graphe sont dessinées à partir de ce nœud vers les nœuds voisins. Les nœuds représentant les unités dont la valeur n'est pas nulle sont de couleur foncée (rouges). La figure 14 montre un échantillon aléatoire de nœuds devant être utilisé comme échantillon initial d'un plan de sondage adaptatif « en toile ». L'échantillonnage adaptatif « se poursuit jusqu'à ce que la taille d'échantillon final visée de 20 unités soit obtenue à la figure 15. À la figure 16, l'échantillon est reconverti aux conditions spatiales. Contrairement à l'échantillonnage adaptatif par grappes, il n'a pas été nécessaire de poursuivre l'échantillonnage jusqu'à ce que chaque unité contenue dans une composante connectée échantillonnée soit incluse. En outre, la faible probabilité d'un saut aléatoire empêche que le plan soit bloqué dans n'importe quelle composante connectée.

Lorsqu'une l'unité i a été sélectionnée, il est possible d'ajouter une étape d'acceptation/rejet pour décider s'il faut l'inclure dans l'ensemble actif, par exemple, l'acceptation avec une probabilité plus élevée si l'unité i possède une valeur élevée ou un degré élevé.

Dans le plan, la constante b proprement dite peut également être remplacée par une probabilité $b(k, t, a_k, y_{a_k}, w_{a_k})$ qui dépend des valeurs reliées aux nœuds et aux liens dans l'ensemble actif ou qui évolue à mesure que progresse la sélection de l'échantillon. Par exemple, si les valeurs des unités comprises dans a_k sont particulièrement élevées, nous pourrions augmenter la probabilité de suivre les liens. Pour ce qui est de la dépendance de b à l'égard de (k, t) , l'utilisation d'un échantillon conventionnel initial de taille $n_0 > 1$ peut être considérée comme servant à obtenir certains renseignements d'après la couverture élémentaire de la population, avant de permettre que l'échantillonnage adaptatif débute.

3. Échantillonnage adaptatif « en toile » spatial

Les plans d'échantillonnage adaptatif tels que l'échantillonnage adaptatif par grappes (Thompson 1990) ont été élaborés en vue de répondre au besoin de stratégies plus efficaces d'échantillonnage de populations spatialement non uniformes, particulièrement celles ayant une distribution géographique rare, par grappes. La plupart des populations possédant une structure de réseau ont aussi une structure géographique ou spatiale inhérente. Par exemple, les populations humaines ont une structure de réseau social, mais sont aussi réparties dans l'espace. Du point de vue du plan d'échantillonnage, un aspect particulièrement intéressant est que les structures spatiales peuvent être caractérisées par des structures en graphes ou en réseau. Par exemple, les relations de quartier fondées sur la proximité géographique peuvent être redéfinies sous la forme de graphes de type treillis. De cette façon, les plans d'échantillonnage par réseau tels que ceux décrits à la section précédente peuvent être appliqués pour résoudre les problèmes d'échantillonnage spatial.

À la présente section, nous décrivons l'utilisation des plans d'échantillonnage adaptatif « en toile » pour échantillonner une population spatialement non uniforme. Ces plans peuvent être considérés comme une généralisation de l'échantillonnage adaptatif par grappes. Sous cet angle, l'échantillonnage adaptatif par grappes serait un cas particulier dans lequel chaque lien est suivi jusqu'à ce qu'il n'y ait plus aucun lien partant de l'échantillon courant. La classe des plans d'échantillonnage adaptatif « en toile » offre toutefois plus de souplesse et de contrôle, et pourrait être plus efficace pour de nombreuses populations spatiales.

$$q_{kit} = b \frac{w_{a_{kt}}}{w_{a_{it}}} + (1 - b) \frac{N}{1 - n_{s_{kt}}} \quad (1)$$

où b est compris entre 0 et 1. S'il n'existe aucun lien partant de l'ensemble actif courant, alors

$$q_{kit} = \frac{N}{1 - n_{s_{kt}}}.$$

Donc, avec la probabilité b , on procède au dépistage de liens, et l'un des liens partant de l'ensemble actif courant est choisi au hasard, ou avec une probabilité proportionnelle à son poids, et le nœud vers lequel il se dirige est ajouté à l'échantillon, tandis qu'avec la probabilité $1 - b$, la nouvelle unité d'échantillon est sélectionnée entièrement au hasard à partir des unités qui n'ont pas encore été sélectionnées. Cependant, s'il n'existe aucun lien ou poids positif partant de l'ensemble actif vers une unité non échantillonnée, l'unité suivante est sélectionnée parmi l'ensemble d'unités non échantillonnées.

L'échantillonnage adaptatif « en toile » de base peut être généralisé de manière à utiliser des liens pondérés. Si la variable de relation w correspond à des poids, au lieu de prendre simplement les valeurs 0 ou 1, la sélection fondée sur les liens peut dépendre de ces poids. Par exemple, les poids de lien peuvent être définis en fonction de la valeur y d'un nœud d'origine ou comme une mesure de distance par rapport au nœud connecté, de sorte que les liens sont suivis avec une plus forte probabilité à partir des nœuds possédant les valeurs les plus élevées ou avec une probabilité plus faible vers les nœuds éloignés. Alors, un lien provenant de l'ensemble actif peut être sélectionné avec une probabilité proportionnelle au poids du lien, ou avec une certaine autre probabilité de sélection $p(i | s_{ckt}, a_k, y_{a_k}, w_{a_k})$ qui ne dépend des variables d'intérêt que par l'intermédiaire de l'ensemble de nœuds actifs. Par exemple, un lien sortant pourrait être sélectionné au hasard parmi les liens dont la valeur de w_{ij} est plus grande qu'une constante donnée ou la probabilité de sélection quand les liens ne sont pas suivis n'a pas à être uniforme sur les unités qui ne figurent pas dans l'échantillon courant, mais peut correspondre à un plan plus général $p(i | s_{ckt})$ tel que la sélection avec une probabilité reliée à une variable auxiliaire ou à partir d'une distribution spatialement définie.

Dans le cas des liens pondérés, w représente une variable de pondération de lien éventuellement continue et la probabilité que cette unité i soit la prochaine unité sélectionnée est

$$q_{kit} = b p(i | s_{ckt}, a_k, y_{a_k}, w_{a_k}) + (1 - b) p(i | s_{ckt}).$$

S'il n'existe aucun lien ou poids positif partant de a_k vers i , alors

des tailles d'échantillon allant jusqu'à dix environ, pas plus que quelques millions de permutations ne devant être énumérées. Pour les plus grandes tailles d'échantillon, les nombres de permutations ou de combinaisons de séquences de sélection possibles dans l'espace des échantillons conditionnels devient trop grand pour effectuer le calcul énumératif exact. C'est pourquoi une approche de rééchantillonnage par chaîne de Markov a été utilisée dans Thompson (2006b) pour calculer les estimateurs améliorés.

La procédure de rééchantillonnage est la suivante. L'objet est d'obtenir une chaîne de Markov x_0, x_1, x_2, \dots ayant une distribution stationnaire $p(x|d_j)$. Ici, x_k désigne un réordonnement complet de l'échantillon à l'étape k de la chaîne. Supposons qu'à l'étape $k-1$, la valeur soit $x_{k-1} = j$, de sorte que h désigne la permutation courante des données d'échantillon dans la chaîne. Une permutation provisoire ou candidate c_k est produite en appliquant le plan d'échantillonnage original, avec la taille d'échantillon n , aux données comme si l'échantillon comprenait l'ensemble de la population, c'est-à-dire comme si $N = n$. Cette distribution de rééchantillonnage, notée p_c , diffère du plan d'échantillonnage réel p , mais possède une certaine similarité avec celui-ci. La distribution conditionnelle souhaitée $p(x|d_j)$ est proportionnelle à la distribution inconditionnelle $p(x)$ sous le plan original appliqué à l'ensemble de la population.

Soit

$$\alpha = \min \left\{ \frac{\bar{p}(c_k)}{\bar{p}(x_{k-1})} \frac{d_c(x_{k-1})}{d_c(c_k)}, 1 \right\}.$$

Avec la probabilité α , t_k est accepté et $x_k = c_k$, tandis qu'avec la probabilité $1 - \alpha$, c_k est rejeté et $x_k = x_{k-1}$.

Cette procédure produit une chaîne de Markov x_0, x_1, x_2, \dots ayant la distribution stationnaire souhaitée $p(x|d_j)$. La chaîne a pour point de départ l'échantillon original s dans l'ordre où il est effectivement sélectionné. Étant donné toute valeur de la statistique exhaustive minimale d_j , la chaîne débute donc dans sa distribution stationnaire et demeure dans cette dernière étape après étape. Supposons que n_j permutations rééchantillonnées soient sélectionnées par ce processus, et soit \hat{p}_{0h} la valeur de l'estimateur initial pour la h^{e} permutation. Un estimateur énumératif de la forme $\hat{p} = E(\hat{p}_{0h} | d_j)$ est remplacé par l'estimateur par rééchantillonnage

$$\hat{p} = \frac{1}{n_j} \sum_{h=1}^{n_j} \hat{p}_{0h}.$$

L'inférence bayésienne fondée sur un modèle avec les plans d'échantillonnage adaptés « en toile » nécessite aussi l'utilisation de méthodes Monte Carlo par chaîne de Markov (MCMC), sauf dans certaines situations comportant un plan assez simple (Chow et Thompson 2003), dans

2.4 Modification des procédures d'échantillonnage adaptées « en toile »

Les plans d'échantillonnage adaptés « en toile » sont une réalisation des plans de sondage à marche aléatoire. Les plans généraux de ces plans ne possèdent pas les propriétés de distribution stationnaire exactes des plans de marche, puisque plus d'un lien peut être suivi en partant de n'importe quel nœud, que les liens peuvent être suivis à partir d'autres nœuds échantillonnés que celui sélectionné le plus récemment, et que l'échantillonnage peut être effectué sans remise. Cependant, les propriétés de distribution stationnaire d'une marche aléatoire ou d'autres plans de sondage par chaîne de Markov peuvent servir de guide pour approximer les propriétés que l'on attendrait d'un plan d'échantillonnage adapté « en toile » similaire.

Durant l'échantillonnage, au moment de la sélection de la t^{e} unité durant la k^{e} vague, soit w_{akt+} le nombre total de liens sortants, ou le total des valeurs des poids, provenant de l'ensemble de nœuds actifs a_k vers des unités qui ne se trouvent pas dans l'échantillon courant $s^{c_{kt}}$. Autrement dit, $w_{akt+} = \sum_{(i \in a_k, j \in s^{c_{kt}})} w_{ij}$. Quand w est une variable indicatrice, w_{akt+} est le total du nombre net de degrés sortants des unités individuelles dans l'ensemble actif a_k , où le degré sortant net est le degré sortant d'une unité moins le nombre de ses liens vers d'autres unités déjà dans l'échantillon courant.

Pour chaque unité i comprise dans l'échantillon, on enregistre la variable d'intérêt y_i et le degré sortant (ou poids sortant) w_{it+} . En plus, pour chaque paire d'unités (i, j) pour laquelle i ainsi que j figurent dans l'échantillon, les valeurs des variables de liens w_{ij} et w_{ji} sont observées. Considérons comme candidate pour la t^{e} sélection durant la k^{e} vague, une unité i qui ne se trouve pas dans l'échantillon courant, de sorte que $i \notin s^{c_{kt}}$. Supposons que l'ensemble actif courant a_k contient une ou plusieurs unités possédant des liens ou des poids positifs partant vers l'unité i , et soit $w_{ait+} = \sum_{j \in a_k} w_{ij}$ leur total. La probabilité que l'unité i soit l'unité suivante sélectionnée est

2.2.1 Méthodes d'inférence

Les méthodes d'estimation sans biais sous le plan et convergentes sous le plan utilisables avec les plans d'échantillonnage adaptatif « en toile » sont décrits dans Thompson (2006b). Les méthodes d'estimation bayésiennes fondées sur un modèle utilisable avec l'échantillonnage adaptatif « en toile » sont décrites dans Kwanasai (2005). Les estimateurs fondés sur le plan de sondage sont construits en partant d'un estimateur relativement facile à calculer qui dépend de l'ordre de sélection de l'échantillon. Cet estimateur initial est ensuite amélioré par la méthode de Rao-Blackwell, c'est-à-dire en obtenant la valeur prévue de l'estimateur initial conditionnellement à la statistique exhaustive minimale.

2.3 Estimateur fondé sur la moyenne de l'échantillon initial

Supposons que μ_0 est un estimateur sans biais de la moyenne de population qui dépend de l'ordre dans lequel l'échantillon est sélectionné. Si l'échantillon initial de nœuds a été sélectionné par échantillonnage aléatoire simple, un exemple d'estimateur initial sans biais qui dépend de l'ordre est la moyenne de l'échantillon initial. L'estimateur amélioré est de la forme

$$\hat{\mu} = E(\hat{\mu}_0 | d_r) = \sum_{\{s: r(s)=s\}} \hat{\mu}_0(s) p(s | d_r).$$

Ici, s désigne l'échantillon dans l'ordre de sélection, r est la fonction de réduction qui réduit l'échantillon ordonné à s , l'échantillon non ordonné de la statistique exhaustive minimale. Les données réduites d_r consistent en l'échantillon non ordonné ainsi que les valeurs associées de la variable d'intérêt. L'estimateur amélioré $\hat{\mu}$ est la valeur prévue de l'estimateur initial sur l'ensemble des $n!$ réordonnements des données échantillonnées. En calculant l'espérance, chacun des réordonnements est pondéré par la probabilité de sélection $p(s | d_r)$.

D'autres estimateurs initiaux utilisés dans l'échantillonnage adaptatif « en toile » utilisent l'entière des données d'échantillon, mais dépendent de l'ordre et sont fondés sur l'utilisation des probabilités conditionnelles de sélection de chaque nouvelle unité dans l'ordre, sachant les unités sélectionnées antérieurement. Quatre types d'estimateurs fondés sur le plan utilisables avec l'échantillonnage adaptatif « en toile » sont donnés dans Thompson (2006b). Le calcul de l'estimateur amélioré $\hat{\mu}$ et de ses estimateurs de variance. Le calcul direct est rapide et efficace pour calculée, ainsi que les valeurs des estimateurs et des estimateurs de variance. Le calcul direct est rapide et efficace pour

Figure 9 Le nœud suivant est sélectionné en suivant l'un des liens partant de l'échantillon courant

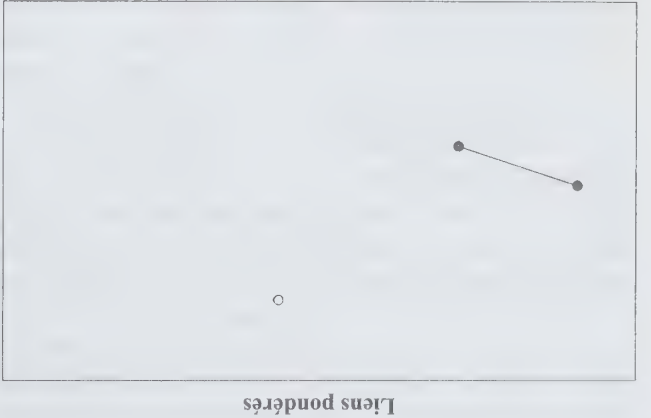


Figure 10 La sélection suivante n'est pas faite par la voie d'un lien partant du dernier nœud sélectionné, mais d'un nœud antérieur

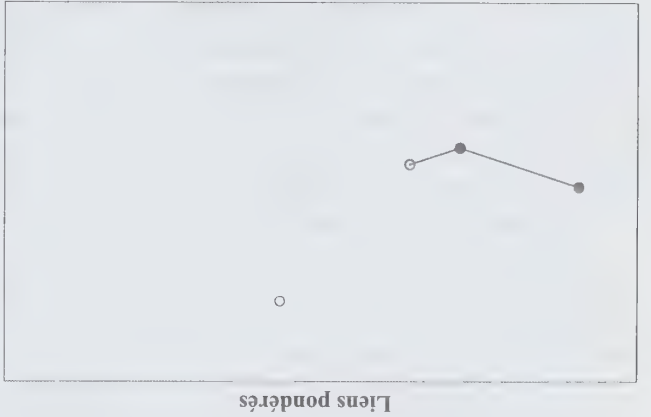
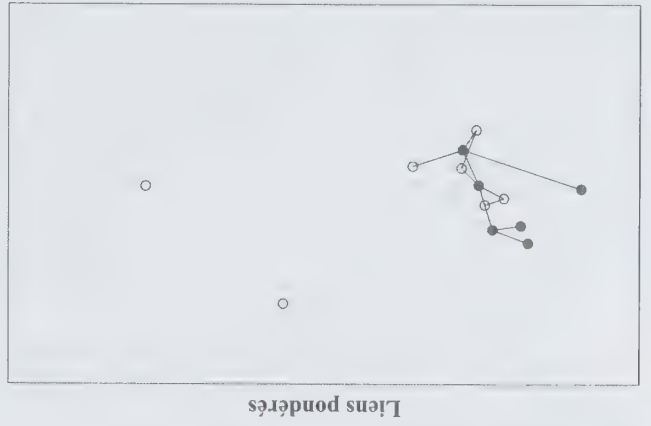


Figure 11 À mesure qu'il progresse, l'échantillonnage est libre de se ramifier souplement dans diverses directions, ainsi que de sélectionner de nouveaux nœuds au hasard dans la population



partir de la population complète est permise. Les plans d'échantillonnage peuvent être appliqués avec ou sans remise.

De manière plus générale, un ensemble de liens peuvent être sélectionnés à chaque étape. En outre, à chaque étape, les liens peuvent être sélectionnés selon un plan plus complexe que l'échantillonnage aléatoire simple. Les probabilités de sélection peuvent dépendre des caractéristiques du nœud ou du lien, et peuvent varier au cours du temps.

L'idée fondamentale qui sous-tend un plan d'échantillonnage adaptatif « en toile » est illustrée par l'ensemble de figures qui suit. À la figure 8, un premier échantillon de deux nœuds a été sélectionné par échantillonnage aléatoire sans remise. À l'étape suivante, on peut choisir au hasard un lien partant de l'un ou l'autre des nœuds initiaux pour ajouter un nouveau nœud à l'échantillon, comme l'illustre la figure 9. Le nœud suivant est sélectionné en suivant un des liens partant de l'échantillon courant. Dans le cas d'une marche aléatoire, un lien devrait être suivi à partir du dernier nœud sélectionné, mais dans le cas de l'échantillonnage adaptatif « en toile », tout lien admissible partant de l'échantillon courant (ensemble de nœuds actifs) peut être suivi. Notons que la sélection suivante, illustrée à la figure 10, n'est pas faite par la voie du lien issu du nœud sélectionné le plus récemment, mais d'un lien provenant d'un nœud antérieur. À mesure qu'il progresse, l'échantillonnage est libre de se ramifier supplémentairement dans différentes directions, de même que de sélectionner de nouveaux nœuds au hasard parmi la population (figure 11). Le plan d'échantillonnage peut être arrêté lorsqu'on atteint une taille d'échantillon particulière ou un autre critère spécifique. Dans le plan illustré par les figures, les liens partant de l'échantillon courant n'ont pas été sélectionnés entièrement au hasard, mais en accordant une plus grande probabilité de suivre les liens partant d'individus à haut risque, représentés par les nœuds foncés ou rouges. En outre, le plan illustre comment une probabilité de 0,1 de sélectionner le nouveau nœud au hasard à n'importe quelle étape au lieu de suivre un lien.

Liens pondérés



Figure 8 Les deux premiers nœuds sélectionnés au hasard

Dans les situations pratiques pour lesquelles nous essayons de trouver des modèles et des plans de sondage appropriés, les poids peuvent être donnés au moins partiellement par les circonstances naturelles de la situation. Par exemple, le poids w_j peut représenter la présence ou l'absence d'un lien allant de la personne i à la personne j , ou le nombre de transactions d'un certain type durant une période donnée allant de i à j . Dans ce cas, la condition (2a) susmentionnée ne serait généralement pas satisfaite et la condition (2b) ne serait satisfaite que si tous les poids étaient symétriques, c'est-à-dire si $w_j = w_i$ pour tout i et j .

En particulier, si certains ou tous les poids sont asymétriques, avec $w_j \neq w_i$, alors (2a) ne serait habituellement pas satisfaite et il ne serait pas possible de choisir arbitrairement les poids pour imposer la condition, parce que généralement les probabilités stationnaires ne seraient pas connues et ne pourraient pas être calculées d'après les données d'échantillon. Cependant, bien que les totaux de ligne w_i ne puissent pas être imposés arbitrairement, ils peuvent être connus pour les unités comprises dans l'échantillon, puisqu'ils sont simplement égaux au poids total des liens partant de chaque unité.

2.2 Échantillonnage adaptatif « en toile »

Les plans d'échantillonnage à marche aléatoire ciblée offrent beaucoup plus de souplesse et de contrôle que les marches aléatoires ordinaires. L'utilisation de liens pondérés avec ces plans de sondage les rend encore plus souples. Cette souplesse demeure néanmoins limitée par la contrainte voulant que la sélection du prochain lien à suivre ne puisse dépendre que du nœud sélectionné le plus récemment dans l'échantillon. L'élaboration de l'ensemble de plans que nous allons décrire avait pour motif d'éliminer cette contrainte et d'élargir considérablement la portée de la souplesse et du contrôle dans les stratégies d'échantillonnage disponibles.

Dans un plan d'échantillonnage adaptatif « en toile » (Thompson 2006b), un premier échantillon d'une ou plusieurs unités (ou nœuds) est sélectionné selon un plan d'échantillonnage aléatoire simple ou un autre plan d'échantillonnage classique. À partir de là, à chaque étape de l'échantillonnage, il existe un ensemble d'unités actives correspondant à l'échantillon sélectionné jusqu'à l'étape en question ou un sous-ensemble de cet échantillon. Dans le cas le plus simple, un lien est sélectionné parmi les liens partant de cet ensemble. L'échantillonnage se poursuit de cette façon jusqu'à ce que la taille souhaitée d'échantillon soit obtenue ou qu'un autre critère d'arrêt soit satisfait. Néanmoins, une faible probabilité que le nœud suivant soit sélectionné au hasard, ou selon un autre plan classique, à

Si les poids ne sont pas symétriques, les probabilités de sélection du plan de sondage à marche aléatoire s'approcheront encore d'une distribution stationnaire à condition qu'il n'existe qu'une seule composante ou, dans la négative, que des sauts aléatoires soient intégrés dans le plan. Cependant, avec les poids pondérés directionnels, la distribution stationnaire n'a plus la forme simple qui peut être calculée d'après les données d'échantillon.

2.1.3 Différentes utilisations des plans de sondage avec liens pondérés

Certaines variantes des plans de sondage avec liens pondérés pourraient s'avérer utiles dans le genre de situations qui suivent.

- 1) Plans de sondage utilisant les poids généraux des liens, sur une échelle continue ou discrète, représentant la force ou l'importance des relations et la probabilité de les suivre.
- 2) Situations comportant deux types de liens, représentés par deux poids, tels que des réseaux sociaux avec des liens relationnels forts et faibles, ou une étude des personnes courant le risque d'une infection par le VIH axée sur les contacts sexuels et les relations d'utilisation de drogue.
- 3) Conditions d'enquête dans lesquelles les liens représentent la partie géographique ou à « sauts aléatoires » du plan, ou le plan d'amorçage. Par exemple, toutes les personnes comprises dans une strate géographique donnée sont unies par un lien géographique, ou toutes les personnes qui visitent l'un des lieux sur une carte ethnographique sont reliées de cette façon.
- 4) Dans une situation où existe une base de sondage, mais que celle-ci ne couvre qu'une partie de la population, toutes les unités comprises dans la base de sondage peuvent être considérées comme étant reliées par un « lien de base de sondage ». L'échantillonnage fondé sur les lieux de rencontre est un exemple type de ce genre de situation.
- 5) L'utilisation d'une variante du plan d'échantillonnage comme modèle de la façon dont un virus ou un autre agent infectieux « échantillonne » les membres d'une population. Un type de plan avec lien pondéré pourrait être élaboré comme modèle pour la propagation d'une maladie infectieuse, en trouvant l'importance différentielle des divers liens. Dans le cas de la grippe, l'importance relative des gouttelettes transportées par l'air (éternuement, toux) par opposition au contact indirect avec des objets solides (poignées de porte, argent). Dans le cas de l'infection par le VIH, l'importance relative des divers types de contacts sexuels et des injections dangereuses,

2.1.4 Propriétés des plans de sondage avec liens pondérés et des graphes de population connexes

qu'il s'agisse de drogues illicites ou d'injection de médicaments dans des conditions insalubres, particulièrement dans les pays du tiers-monde. La transmission de la maladie dans un contexte de simulation suit un protocole légèrement différent de celui des plans de sondage mis en œuvre, en ce sens qu'au lieu de penser à un nouveau lien sélectionné à chaque étape temporelle de sélection, il pourrait exister un nombre de transmissions variant de zéro à un chiffre élevé durant une étape temporelle.

Supposons que des poids soient affectés aux relations au sein de la population, le poids w_{ij} désignant la force de la relation allant du nœud i au nœud j . Supposons aussi que nous utilisons un plan par dépistage de liens de type marche dans laquelle la probabilité de transition est

$$P_{ij} = \frac{w_{ij}}{w_i}$$

où $w_i = \sum_{j=1}^N w_{ij}$. Il s'agit de la probabilité conditionnelle de sélectionner le nœud j comme unité d'échantillon suivante, sachant que l'unité sélectionnée le plus récemment est le nœud i . Le plan de marche est une chaîne de Markov sur un graphe, dans lequel les liens sont pondérés. Nous allons maintenant considérer la question dans l'autre direction, c'est-à-dire quand une chaîne de Markov peut être représentée par un plan de cette sorte sur un graphe présentant des liens pondérés. Étant donné une chaîne de Markov spécifiée par une matrice de probabilités de transition P_{ij} , nous pouvons toujours la représenter comme un plan de sondage à marche de ce type sur un graphe dont les liens sont pondérés, à condition que les liens satisfassent la première des propriétés suivantes :

- 1) $w_{ij} = P_{ij} w_i$, où les totaux des poids de ligne sont choisis arbitrairement.
- Ensuite, considérons l'imposition d'une certaine propriété aux totaux des poids de ligne afin de les rendre uniques. Par exemple :
- 2a) Si les totaux des poids de ligne w_i sont choisis de manière qu'ils soient tous égaux, ou qu'ils soient tous égaux à une constante telle que un, alors les poids des liens représentent les probabilités de transition conditionnelles, sachant que le processus se trouve au nœud d où ils sont originaires.
- 2b) Si les totaux des poids de ligne w_i sont proportionnels aux probabilités stationnaires π_i de la chaîne de Markov pour chaque nœud i , ou qu'ils y sont égaux, les poids représentent les « flux » de la chaîne de Markov, c'est-à-dire les probabilités in-

conditionnelles des transitions le long des liens :

approximation assez proche des propriétés des stratégies ciblées dans les comparaisons empiriques (Thompson 2006a).

2.1.1 Plans de sondage avec des liens pondérés

Dans de nombreuses études portant sur des populations formant un réseau social, le réseau est conceptualisé comme étant constitué de nœuds (personnes) et de lignes ou de flèches représentant les liens ou les relations entre les personnes. Le réseau est caractérisé par une matrice d'incidence contenant des 0 et des 1 indiquant quand il existe un lien allant du nœud (ligne) i au nœud (colonne) j . Cependant, dans de nombreuses situations réelles, on peut s'intéresser à plus d'un type de lien et les liens peuvent posséder différents poids représentant différentes forces d'une relation. Par exemple, dans les études des comportements à risque et des interventions reliées à l'épidémie d'infection par le VIH, deux types de liens auxquels on accorde beaucoup d'intérêt sont les relations sexuelles et les relations de consommation de drogues injectables. D'autres relations sociales, telles que les amitiés et les modalités de logement, peuvent aussi intéresser les chercheurs et être utiles pour trouver les membres de la population. On peut donner à ces types de relations des poids qui correspondent à la fréquence des rencontres, à la proximité géographique ou à d'autres mesures de force.

Dans la forme élémentaire de plan de sondage s'appuyant sur des liens pondérés que nous considérons, dans laquelle est choisi l'un des liens partant de la personne sélectionnée le plus récemment, la sélection est faite avec une probabilité proportionnelle au poids du lien. Plus généralement, la sélection pourrait être faite en se basant sur ce poids, mais pas nécessairement de manière proportionnelle à celui-ci. Cependant, nous pourrions alors redéfinir le poids pour qu'il soit proportionnel à la probabilité que nous avons sous le plan de suivre ce lien, de sorte que le résultat qui suit serait encore applicable.

La dérivation qui suit montre que, sous des conditions appropriées, la probabilité de sélection stationnaire de chaque personne lorsque l'on se sert de ce genre de plan est proportionnelle à la somme des poids des liens partant de cette personne. Le résultat s'applique pour une population dans laquelle il est possible de rejoindre n'importe quelle personne à partir d'une autre en suivant une certaine trajectoire dans laquelle chaque lien possède un poids plus grand que zéro. Autrement dit, la population ne comprend qu'une seule composante.

Pour qu'une telle condition soit vérifiée, il est avantageux d'avoir au moins une certaine probabilité de suivre des liens communs, mais faibles. Par exemple, dans une étude d'une épidémie d'une maladie transmissible sexuellement, on pourrait vouloir se concentrer avec une forte probabilité sur

les liens sexuels. Mais ces liens sexuels ne relient pas la population en une seule composante. Par conséquent, une probabilité plus faible est permise dans le plan de sondage afin de suivre les liens d'amitié ou les liens géographiques, qui représentent des relations plus faibles entre les personnes et qui intéressent intrinsèquement moins les chercheurs, mais qui servent à relier la population. Donc, dans cette situation, la combinaison de divers types de liens transforme la population en une composante unique pour les besoins du plan de sondage.

2.1.2 Distribution stationnaire d'un plan de sondage par chaîne de Markov avec liens pondérés

À la présente section, nous déterminons la distribution stationnaire d'un plan de sondage avec liens pondérés dans le cas d'une population à une seule composante. Il ne faut pas perdre de vue que nous pouvons créer la propriété de composante unique grâce à l'emploi innovateur de liens géographiques conjugués à des liens sociaux.

Soit w_{ij} le poids d'un lien entre le nœud i et le nœud j , et supposons que ces liens sont symétriques, de sorte que $w_{ij} = w_{ji}$. Considérons un plan de sondage à marche aléatoire, avec remise, dans lequel la probabilité de transition vers le nœud j , sachant que la marche est au nœud i , est proportionnelle à w_{ij} . Autrement dit, un lien est sélectionné à la sortie du nœud i avec une probabilité proportionnelle au poids. La probabilité de transition est donc $P_{ij} = w_{ij} / w_i$. La somme $w_i = \sum_j w_{ij}$ représente le poids total des liens sortant du nœud i , ce qui généralise le concept de degré de nœud avec des nœuds équi pondérés.

Supposons que le graphe ne possède qu'une seule composante, c'est-à-dire que tout nœud dans le graphe peut être atteint en partant de n'importe quel autre nœud en suivant une trajectoire dans laquelle chaque lien possède un poids positif. Alors, la probabilité stationnaire pour le nœud i est proportionnelle à w_i .

Supposons que la probabilité que la marche soit au nœud i au temps t est $\pi_i = w_i / w$, pour $i = 1, \dots, N$, où $w = \sum_j \sum_i w_{ij}$, le total de tous les poids. Alors, la probabilité que le processus soit au nœud i au temps $t + 1$ est $\sum_j \pi_j P_{ji}$ en vertu de la loi des probabilités totales. En ce qui concerne les poids de lien, cette somme est $\sum_j (w_j / w) (w_{ji} / w_i) = \sum_j w_{ji} / w$. Étant donné la symétrie des liens pondérés, cette expression devient w_i / w , de sorte que, si le nœud i possède cette probabilité au temps t , il a la même probabilité au temps $t + 1$, de sorte qu'il s'agit des probabilités stationnaires du processus. Par induction, une fois que le processus atteint sa distribution stationnaire, il y reste à chaque étape ultérieure. En pratique, surtout si les tailles d'échantillon sont petites ou pour diverses variantes du plan de sondage, la distribution stationnaire sert d'approximation de la distribution exacte.

probabilité de sélection double de celle d'un individu ne présentant pas le comportement en question.

C'est l'échantillon de personnes ou nœuds accepté qui possède les probabilités de sélection stationnaire souhaitées. Si les unités (nœuds) susceptibles d'être sélectionnées étaient interviewées en profondeur également, au lieu de se limiter à l'interview de sélection au sujet du degré sortant du nœud, en principe, les estimations d'après l'échantillon accepté pourraient être améliorées en appliquant la méthode de Rao-Blackwell (Casella et Robert 1996). Cela comprendrait le calcul des probabilités d'obtenir les mêmes données avec différents résultats d'acceptation-rejet et dans différents ordres de sélection. Pour chacun des divers scénarios d'acceptation, l'estimation serait calculée en se servant de l'ensemble accepté et de chaque valeur pondérée par les probabilités de sélection et d'acceptation ordonnées. Dans la plupart des cas, le nombre de combinaisons est trop élevé pour effectuer un calcul exact, et une approche plus pratique serait la méthode de rééchantillonnage par chaîne de Markov à l'étape de l'inférence décrite dans une section ultérieure du présent article. Il n'est pas certain qu'en pratique, il soit souhaitable de calculer les estimateurs améliorés en se servant des données, puisque des interviews complètes plutôt que des interviews de sélection seraient nécessaires pour les personnes qui ne sont pas acceptées au départ, que les calculs pour l'amélioration peuvent être lourds et qu'ils dépendent de la connaissance des probabilités de sélection pour l'échantillon initial, ce qui n'est pas nécessaire pour les estimateurs simples.

Dans le cas d'un plan de sondage à marche ciblée dans lequel la probabilité de sélection stationnaire cible π_i du nœud i est proportionnelle à c_i , un estimateur asymptotiquement convergent, fondé sur les probabilités limites, est fourni par l'estimateur par le ratio généralisé

$$\hat{\pi}_i = \frac{\sum_s y_i / c_i}{\sum_s y_i / c_i}$$

où y_i est la valeur de la variable d'intérêt pour le i^{e} nœud et s_a est l'échantillon de nœuds sélectionnés. Dans ce type d'estimateur, les valeurs relatives des probabilités cibles doivent être spécifiées, puisque la constante de proportionnalité s'annule.

Il convient de souligner qu'on ne peut pas utiliser un simple estimateur de Horvitz-Thompson ou de Hansen-Hurwitz, parce que l'on ne connaît pas la constante de proportionnalité dans les probabilités d'inclusion, tandis que dans l'estimateur par le ratio généralisé, cette constante s'annule. De nouveau, les probabilités limites sur lesquelles l'estimateur est fondé sont vérifiées exactement pour le plan de sondage avec remise. Pour la variante sans remise, les propriétés de l'approche avec remise constituaient une

Puisqu'une marche aléatoire avec remise dans un graphe ou un réseau est une chaîne de Markov, les notions de la méthode Monte Carlo par chaîne de Markov peuvent être appliquées pour produire une chaîne de Markov différente ayant les probabilités stationnaires souhaitées. À chaque étape de l'échantillonnage, l'état de la chaîne est le nœud courant ajouté à l'échantillon. Les probabilités stationnaires de la chaîne correspondent aux probabilités de sélection stationnaire pour chaque personne ou nœud. Dans le cas d'un plan de sondage à marche ciblée, le plan de sondage à marche aléatoire est modifié légèrement à chaque pas, en fonction du degré sortant de chaque nœud, pour obtenir un plan avec les probabilités de sélection limites spécifiées.

Supposons qu'à une étape de l'échantillonnage, la personne i soit la dernière personne qui a été ajoutée à l'échantillon. En suivant une procédure de marche aléatoire, nous sélectionnons aléatoirement l'un des liens partant de cette personne, et ce lien mène à la personne j , qui est maintenant l'unité que nous tentons de sélectionner. Une interview de sélection révèle que la personne j possède plus de liens sortants que la personne i , de sorte que la probabilité conditionnelle d'aller de i à j comme nous venons de le faire est plus grande que la probabilité conditionnelle dans la direction opposée, puisque les probabilités de transition sont reliées à l'inverse du nombre de liens sortants. Par conséquent, nous calculons une probabilité inférieure à l'unité et acceptons la personne j dans l'échantillon uniquement avec cette probabilité. Si notre sélection provisoire n'est pas acceptée, nous choisissons de nouveau de manière indépendante un lien partant de la personne i . La probabilité d'acceptation du lien candidat est fondée sur la généralisation de Hastings (1970) de l'algorithme de Metropolis. La probabilité d'acceptation dépend des probabilités de sélection cibles souhaitées, du nombre de liens sortants du nœud courant et du nœud candidat, ainsi que de la probabilité qu'un saut aléatoire se fasse dans l'une ou l'autre direction si cela fait partie du plan de sondage (Thompson 2006a).

Notons que la méthode dépend uniquement des liens sortants, qui peuvent habituellement être déterminés pour les membres de l'échantillon, alors que les liens arrivant vers les membres de l'échantillon (liens entrants) ne peuvent habituellement pas être déterminés. Par conséquent, la méthode s'applique aux réseaux directionnels ainsi que symétriques.

Un plan à marche uniforme est le cas particulier dans lequel les probabilités de sélection stationnaire ciblées sont toutes égales. Un plan de sondage à marche aléatoire ciblée pourrait être utilisé par exemple pour obtenir un échantillon d'une population cachée dans laquelle un individu présentant un certain comportement à haut risque aurait une

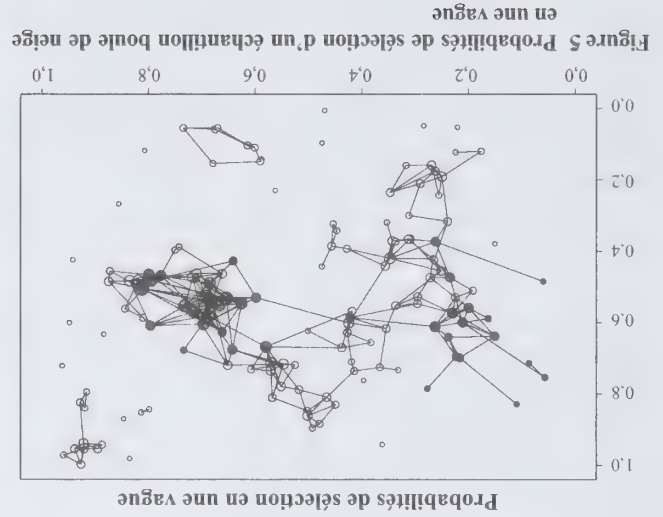
d'une vague, il n'est généralement pas possible de calculer les probabilités d'inclusion des nœuds d'après les données d'échantillon. Des méthodes de calcul d'estimateurs sans biais sous le plan des caractéristiques de population des nœuds et des liens pour ce genre de plans sont décrites plus loin dans le présent article, à la section sur l'échantillonnage adaptatif « en toile ».

La figure 6 montre un échantillon boule de neige

provenant de la même population en réseau ayant pour point de départ une unité sélectionnée au hasard. Puisque la population est constituée de plus d'une composante connectée, un plan à marche aléatoire strict resterait bloqué dans la composante dans laquelle il a débuté, quelle qu'elle soit. Il est par conséquent souhaitable de fournir dans le plan d'échantillonnage, à chaque étape, une petite probabilité de sélection de l'unité suivante par échantillonnage aléatoire simple ou selon un autre plan d'échantillonnage classique, ou au moins de permettre un saut aléatoire chaque fois que l'on s'aperçoit qu'une marche est bloquée dans une composante.

La figure 7 montre les probabilités de sélection stationnaires pour la marche aléatoire à travers le réseau illustré.

Bien que ces probabilités dans cette population ne soient pas simplement proportionnelles aux degrés de nœud, on peut voir que les nœuds ayant un degré élevé ont tendance à avoir des probabilités élevées de sélection. En outre, puisque les personnes à risque élevé dans cette population ont tendance à avoir une probabilité de sélection élevée sous ce plan de sondage, les sommes d'échantillon, tels que les moyennes d'échantillon et les proportions d'échantillon, ne sont pas des estimateurs sans biais des moyennes et des proportions de population. Pour obtenir des estimations sans biais, il faudrait utiliser les méthodes décrites aux sections suivantes du présent article.



2.1 Plans de sondage à marche aléatoire ciblée

Figure 6 Un échantillon sélectionné par marche aléatoire à partir de la même population

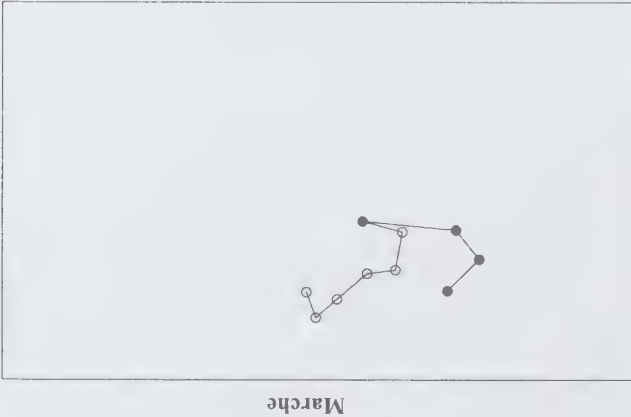
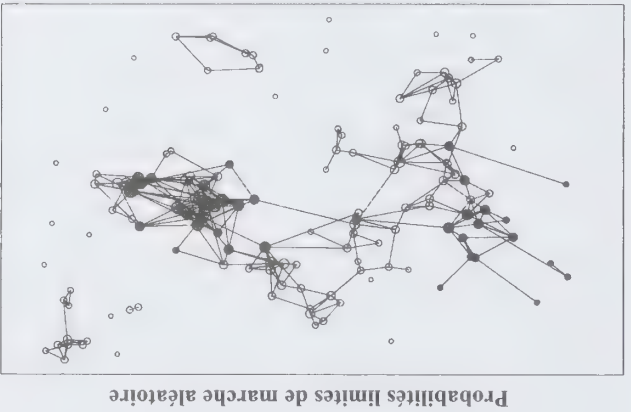


Figure 7 Probabilités de sélection limites de marche aléatoire



L'une des principales raisons d'utiliser des plans de sondage à marche aléatoire pour échantillonner les populations cachées était de pénétrer plus en profondeur dans la population, c'est-à-dire d'aller au-delà de l'échantillon initial et donc d'obtenir un échantillon plus représentatif de la population. Quand les probabilités de sélectionner une personne particulière par une méthode de ce genre sont calculées étape par étape ou dans leur limite stationnaire, elles ne sont en général pas égales, mais dépendent de la structure des liens et des degrés de nœud. Des plans de sondage à marche aléatoire uniforme et ciblée ont été élaborés en premier lieu en vue de trouver une méthode de tirage d'un échantillon au moyen d'un réseau telle que les probabilités stationnaires soient les mêmes pour chaque personne ou nœud (Thompson 2006a). Une motivation supplémentaire était de trouver un moyen plus ciblé et adaptable d'effectuer un échantillonnage par réseau.

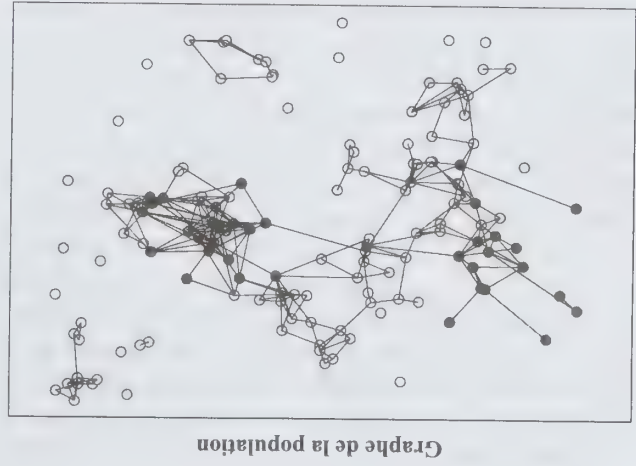


Figure 1 Une population ayant une structure en réseau

La figure 2 montre un échantillon aléatoire simple initial de cinq nœuds sélectionnés dans la population structurée en réseau. Un échantillon boule de neige à une vague sélectionné en suivant chaque lien sortant de l'échantillon initial est illustré à la figure 3, et un échantillon boule de neige à deux vagues issu du même échantillon initial est illustré à la figure 4. Notons qu'avec un nombre fixe de vagues, un échantillon boule de neige peut croître très rapidement.

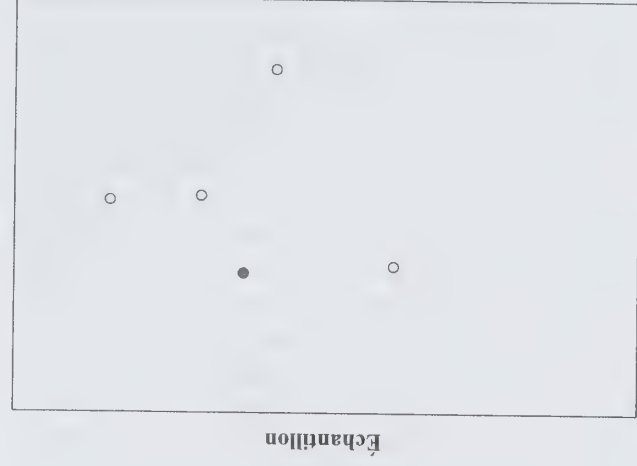


Figure 2 Un échantillon aléatoire de nœuds

Dans le cas des plans d'échantillonnage boule de neige et de nombreux autres plans d'échantillonnage par dépistage de liens, les sommes des données d'échantillon, tels qu'une moyenne d'échantillon ou des proportions d'échantillon, ne sont pas de bons estimateurs des caractéristiques de population analogues. En effet, sous le plan, différentes unités ont des probabilités différentes de sélection, qui dépendent de la structure des liens dans la population. La

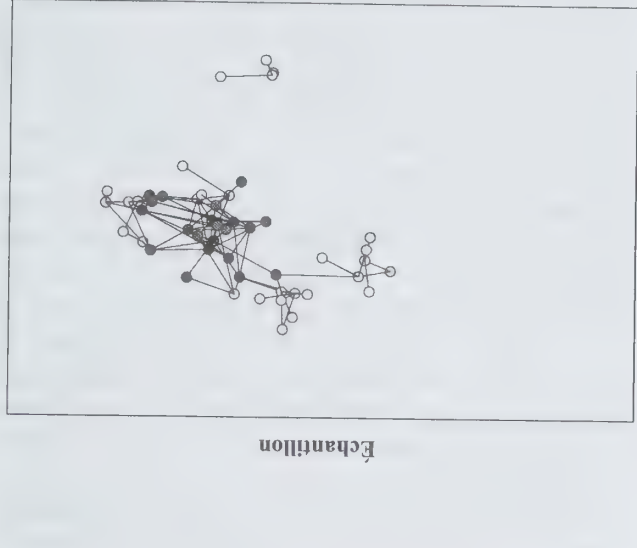


Figure 3 Échantillon boule de neige à une vague

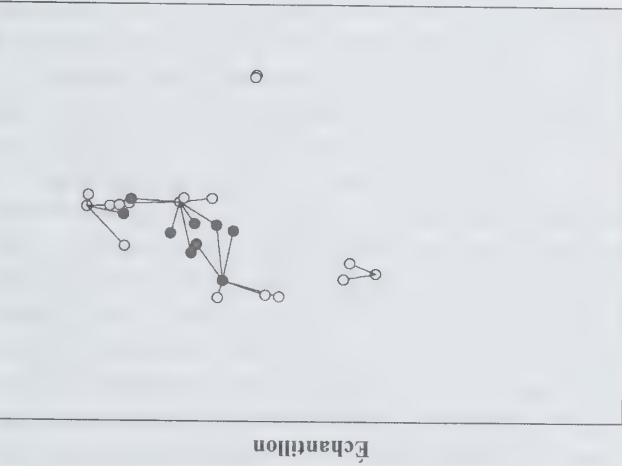


Figure 4 Échantillon boule de neige à deux vagues

Dans le cas du plan d'échantillonnage boule de neige à une vague dans des conditions où les liens sont symétriques, les probabilités d'inclusion des nœuds d'échantillon peuvent être calculées facilement comme étant proportionnelles au degré des nœuds. Si les liens sont asymétriques ou que les plans d'échantillonnage boule de neige comprennent plus

utilise le concept d'une distribution stationnaire d'une marche aléatoire dans un graphe pour développer un moteur de recherche et un algorithme de classement des pages Web, en évoquant la métaphore d'un « intermédiaire aléatoire » pour décrire le processus d'une marche aléatoire qui suit des hyperliens d'une page Web à l'autre.

Heckathorn (1997, 2002), ainsi que Saiganiak et Heckathorn (2004) ont décrit une méthode d'échantillonnage appelée « échantillonnage déterminé selon les répondants » dans le cadre de laquelle on se servait d'un système de coupons pour motiver les membres d'une population cachée à recruter d'autres membres de la population dans l'échantillon. Un estimateur simple des totaux et des moyennes de population, dans lequel chaque observation est pondérée par l'inverse du degré de nœud de la personne en question a été utilisé dans ces plans de sondage fondés sur la distribution limite d'une marche aléatoire avec remise dans un réseau présentant des liens symétriques et une seule composante connectée. Les méthodes basées sur un système de coupons élaborées pour ces plans de sondage se sont avérées extrêmement efficaces pour le recrutement d'échantillons d'une taille importante auprès de populations cachées dans un certain nombre de circonstances.

La notation utilisée pour l'échantillonnage par réseau est la suivante. Il existe une population d'unités ou de nœuds étiquetés $1, 2, \dots, N$ à laquelle sont associées les variables d'intérêt y_1, y_2, \dots, y_N . À chaque paire de nœuds (i, j) est associé un indicateur de lien ou poids, de sorte que la série $\{w_{ij}; i, j = 1, \dots, N\}$ contient les variables d'intérêt associées aux paires de nœuds.

Dans le contexte d'un réseau, un échantillon s est un sous-ensemble $s^{(1)}$ de nœuds et un sous-ensemble $s^{(2)}$ de paires de nœuds, c'est-à-dire $s = (s^{(1)}, s^{(2)})$. Donc, l'échantillon est constitué d'un échantillon de nœuds pour lesquels les variables de nœud y d'intérêt sont enregistrées, et d'un échantillon de paires de nœuds, pour lesquelles les valeurs des variables de relation w sont enregistrées.

La figure 1 montre une population structurée en réseau que nous utiliserons pour illustrer certains plans d'échantillonnage par réseau décrits dans le présent article. Dans le cas d'une population humaine ayant une structure de réseau social, les nœuds rouges ou de couleur foncée pourraient représenter des personnes pour lesquelles les valeurs des variables d'intérêt sont élevées, par exemple indiquant un comportement à risque, tel que la prise de drogues injectables. Les nœuds jaunes ou de couleur pâle représenteraient les personnes ne présentant pas la caractéristique à haut risque d'intérêt. Les liens entre les personnes représenteraient les relations sociales, telles que la prise de repas ensemble, les relations relatives à l'utilisation de la drogue, ou les contacts sexuels.

calculs, est l'estimateur fondé sur la multiplicité qui consiste simplement à diviser la valeur observée d'une variable d'intérêt mesurée sur une unité d'observation par la « multiplicité » de cette unité, c'est-à-dire le nombre d'unités de sélection à laquelle elle est liée. Des estimateurs de Horvitz-Thompson ont également été introduits pour la stratégie. Au cours des décennies suivantes, de nombreuses variantes de cette stratégie ont été publiées dans la littérature statistique sur le sujet.

Dans l'échantillonnage boule de neige, un échantillon initial de nœuds est sélectionné suivant un plan donné, tel que l'échantillonnage aléatoire simple, et chaque lien partant de ces nœuds est suivi pour ajouter à l'échantillon les nœuds reliés. Ce processus est poursuivi pendant un nombre spécifié d'étapes, ou « vagues ». Plus généralement, un sous-échantillon, tel qu'un nombre fixe de liens, est suivi à chaque vague. Frank (1971, 1977a, b, 1978a, b, 1979) a défini le problème comme étant celui d'un échantillonnage par graphes et a élaboré des estimateurs fondés sur le plan pour de nombreux cas d'échantillonnages boule de neige, y compris les plans avec probabilités de sélection initiales égales, et des estimateurs des quantités de population tels que les totaux et les moyennes des variables associées avec des nœuds ou des individus, ainsi que des grandeurs de lien au niveau de la population, telles que le degré de nœud moyen, où le degré d'un nœud est défini comme étant le nombre de liens sortant de ce nœud (ou entrant dans le nœud). Frank et Snijders (1994) ont introduit un certain nombre d'estimateurs sous le plan et sous un modèle pour des plans d'échantillonnage boule de neige à une vague dont l'élaboration avait été motivée par le problème d'estimation du nombre d'utilisateurs de drogues injectables dans une ville.

Dans un plan d'échantillonnage à marche aléatoire, un premier nœud est sélectionné au hasard. Parmi les liens sortants de ce nœud, un lien est sélectionné au hasard et suivi pour ajouter à l'échantillon le nœud connecté. Ce processus se poursuit pendant un nombre spécifié de vagues, une unité étant sélectionnée à chaque vague. Si l'échantillonnage est effectué avec remise, le plan est une chaîne de Markov dont l'état à chaque étape est l'identité du nœud sélectionné à l'étape en question. Les propriétés des plans de ce genre, représentés comme des chaînes de Markov dans des graphes, telles que les probabilités limites ou les probabilités stationnaires ont été examinées dans la littérature sur la statistique et les probabilités (Lovász 1993). Les plans d'échantillonnage à marche aléatoire ont été introduits dans la littérature sur les réseaux sociaux par Klovadahl (1989) en vue de pénétrer dans une population humaine cachée plus au-delà de l'échantillon initial qu'il n'est possible de le faire avec la même taille d'échantillon en utilisant un plan d'échantillonnage boule de neige. Dans la littérature sur l'informatique, Brin et Page (1998) ont

moyen d'un graphe qui consiste en un ensemble de nœuds et un ensemble d'arêtes ou d'arcs entre les nœuds. De façon plus générale, chaque relation entre une paire de nœuds peut posséder un poids reflétant la force de la valeur associée à la relation.

Les populations humaines ont une structure en réseau inhérente découlant des relations sociales. Comme nous le mentionnerons plus loin, les relations spatiales donnent aussi une structure en réseau à de nombreuses populations. Les réseaux informatiques, les réseaux de communication, les réseaux de régulation des gènes et les réseaux métabo-

liques donnent aussi naissance à des populations en réseau. La structure en réseau des populations est importante pour deux raisons. Premièrement, les liens existant dans les réseaux peuvent présenter un intérêt en soi pour les chercheurs. Par exemple, dans le cas d'une épidémie d'une maladie contagieuse, il est important de connaître la nature et le profil des contacts sociaux par lesquels se propage la maladie. Deuxièmement, la structure en réseau peut être utilisée pour faciliter l'obtention d'un échantillon auprès d'une population par ailleurs difficile à échantillonner. Par exemple, dans l'étude des populations cachées courant un risque d'infection par le VIH ou de SIDA, y compris les utilisateurs de drogues injectables, les travailleurs du sexe et d'autres, dans de nombreux cas, le seul moyen d'obtenir un échantillon suffisamment grand pour l'étude consiste à suivre les liens sociaux en partant des personnes échantillonnées au départ pour trouver un plus grand nombre de membres de la population cachée.

La plupart des plans d'échantillonnage par réseau qui consistent à suivre des liens sont intrinsèquement adaptés en ce sens que les valeurs des liens utilisés dans la sélection sont des variables d'intérêt qui ne sont généralement pas connues avant l'enquête. En outre, dans certaines études, il peut être intéressant de suivre les liens ayant une probabilité plus élevée en partant des personnes échantillonnées pour lesquelles les valeurs des variables associées au comportement à risque sont élevées.

Une catégorie de plans appelée *échantillonnage fondé sur la multiplicité* ou simplement *échantillonnage par réseau* a été introduite par Birnbaum et Sitken (1965) en même temps que des estimateurs sans biais sous le plan des quantités de population. L'approche a été approfondie par Sitken (1970, 1972a, b) et d'autres, et est décrite dans Thompson (2002). Dans ces plans, les unités sur lesquelles sont faites les observations sont obtenues en tirant d'abord des « unités de sélection » auxquelles les unités d'observation sont liées. Ces stratégies ont été motivées par des problèmes de santé publique pour lesquels on s'est aperçu que les estimations habituellement utilisées étaient biaisées, à cause du nombre inégal de ce genre de liens. Le plus simple des estimateurs sans biais, pour ce qui est des

haut rendement en ce sens que les valeurs d'échantillon des variables d'intérêt ont tendance à être plus élevées, en moyenne, que les moyennes de population de ces variables. Bien qu'il s'agisse souvent d'une caractéristique souhaitée en soi dans les études de population rare, les simples sommes de données d'échantillon, tels que les moyennes et les proportions d'échantillon ne sont généralement pas de bonnes estimations des moyennes et des proportions de population. À leur place, des estimateurs des quantités de population fondées sur le plan de sondage et fondés sur un modèle ont été élaborés en vue de leur utilisation avec des plans de sondage adaptatifs.

Dans le cas des estimateurs fondés sur le plan de sondage, des propriétés telles que l'absence de biais et la convergence dépendent uniquement de la façon dont l'échantillon est sélectionné et non des hypothèses formulées au sujet de la population. Par contre, les estimateurs fondés sur un modèle, tels que les estimateurs du maximum de vraisemblance ou les estimateurs bayésiens, requièrent l'utilisation d'un modèle statistique, dont les valeurs des paramètres sont habituellement inconnues, décrivant la population d'intérêt. Les estimateurs fondés sur le plan de sondage pour plans adaptatifs sont décrits dans Thompson et Seber (1996), Thompson (2006a, b) et des articles antérieurs.

Les résultats fondamentaux concernant les approches fondées sur un modèle de l'inférence sous des plans de sondage adaptatifs ont été présentés dans Thompson et Seber (1996), qui ont montré que les méthodes fondées sur la vraisemblance, telles que la méthode du maximum de vraisemblance, et l'inférence bayésienne seraient plus efficaces que d'autres approches fondées sur un modèle (par exemple, l'approche de la prédiction linéaire sans biais) dans le cas des plans d'échantillonnage adaptatifs. L'estimation du maximum de vraisemblance et l'approche fondée sur la vraisemblance appliquée de manière plus générale dans les plans par dépistage de liens ont été décrites dans Thompson et Frank (2000). L'estimation bayésienne avec des plans par dépistage de liens a été utilisée dans Chow et Thompson (2003). Une méthode combinant des caractéristiques et des approches fondées sur un modèle et fondées sur le plan de sondage a été suivie dans Felix-Medina et Thompson (2004). L'estimation bayésienne en se servant d'une méthode Monte Carlo par chaîne de Markov (MCMC) conjuguée à des plans d'échantillonnage adaptatif « en toile » est décrite dans Kwanisai (2005, 2006).

2. Échantillonnage adaptatif dans un contexte de réseau

Une population a une structure en réseau s'il existe des liens ou des relations entre les unités de la population. Mathématiquement, ce genre de population est décrite au

Echantillonnage adaptatif par réseau et spatial

Steve Thompson¹

Résumé

Le présent article décrit les progrès récents dans le domaine des stratégies d'échantillonnage adaptatif et présente de nouvelles variantes de ces stratégies. Les progrès récents comprennent les plans d'échantillonnage à marche aléatoire ciblée et l'échantillonnage adaptatif « en toile ». Ces plans conviennent particulièrement bien pour l'échantillonnage par réseau; par exemple pour obtenir un échantillon de personnes appartenant à une population humaine cachée en suivant les liens sociaux partant d'un groupe de personnes échantillonnées pour trouver d'autres membres de la population cachée à ajouter à l'échantillon. Chacun de ces plans peut également être transposé à des conditions spatiales pour produire de nouvelles stratégies d'échantillonnage adaptatif spatial souples, applicables à des populations réparties non uniformément. Les variantes de ces stratégies d'échantillonnage comprennent celles où les liens du réseau ou les liens spatiaux ont des poids inégaux et sont suivis avec des probabilités inégales.

Mots clés : Echantillonnage par réseau ; échantillonnage boule de neige ; marche aléatoire ; chaîne de Markov ; échantillonnage adaptatif « en toile ».

1. Introduction

Un plan d'échantillonnage adaptatif est une procédure de tirage d'échantillon dans laquelle les probabilités de sélection d'un ensemble d'unités de la population dans l'échantillon dépendent des valeurs de la variable d'intérêt observées durant l'enquête. Dans un contexte spatial, un exemple d'échantillonnage adaptatif est celui d'une enquête dans laquelle, chaque fois que l'on observe pour une unité échantillonnée une valeur de la variable d'intérêt inhabituellement élevée ou intéressante pour une autre raison, des unités voisines peuvent être ajoutées à l'échantillon. Dans un contexte de réseau, tel qu'une sous-population humaine caractérisée par un réseau social, un plan d'échantillonnage par dépistage de liens peut être utilisé pour suivre de manière adaptative les liens sociaux des individus échantillonnés en vue de repérer et d'ajouter des membres supplémentaires de la sous-population à l'échantillon.

Dans le cas de l'échantillonnage spatial, l'élaboration de plans d'échantillonnage adaptatifs a été motivée par des problèmes tels que l'estimation de l'abondance d'espèces végétales et animale rares et regroupées, l'évaluation de polluants environnementaux répartis non uniformément, l'étude de sous-populations de personnes géographiquement regroupées. Dans les contextes de réseau, l'élaboration de plans d'échantillonnage adaptatif par réseau a été motivée par des problèmes d'échantillonnage de personnes atteintes de maladies rares, d'échantillonnage de populations cibles, telles que celles courant un risque élevée d'infection par le VIH ou de SIDA, ou d'autres maladies épidémiques, ainsi que l'échantillonnage par la voie de réseaux informatiques et de communications.

Zacks (1969) et Basu (1969) ont reconnu que, dans la plupart des cas, l'échantillonnage optimal serait en principe

1. Steve Thompson, Simon Fraser University. Courriel : Thompson@stat.sfu.ca.

Malgré les premiers résultats théoriques et la motivation découlant des opérations d'enquête, il a fallu plusieurs décennies avant que ne soit généralement reconnue l'importance des plans adaptatifs tant en théorie qu'en pratique. L'importance pratique des stratégies d'échantillonnage adaptatif est devenue évidente lorsque l'on a appliqué la réflexion statistique à la résolution de problèmes de gestion des ressources naturelles et de protection de l'environnement. L'élaboration de plans de sondage par dépistage de liens adaptatifs pour rejoindre des populations humaines cachées a acquis une importance stratégique lorsque il s'agit de résoudre des problèmes comme comprendre et enrayer l'épidémie mondiale d'infection par le VIH. En outre, des questions concernant les dépenses et les efforts requis pour les enquêtes sociales de tout type ont suscité un intérêt nouveau pour les méthodes d'échantillonnage adaptatif.

Les plans d'échantillonnage adaptatifs, tels que ceux décrits dans le présent article, servent souvent de plans à

un échantillonnage adaptatif. Sous un modèle bayésien de la population, à n'importe quelle étape de la procédure d'échantillonnage, au lieu d'un plan d'échantillonnage classique, on peut obtenir des résultats aussi bons, voire meilleurs, en sélectionnant l'échantillon restant de manière à obtenir l'erreur quadratique moyenne la plus faible sachant les valeurs d'échantillon observées jusque-là. L'erreur quadratique moyenne globale est la valeur attendue de l'erreur quadratique sous-jacente est que l'intégrale du minimum d'un ensemble de fonctions est plus petite, ou n'est pas plus grande que le minimum des intégrales. Les résultats concernant les stratégies adaptatives optimales sont décrits et étendus dans Thompson et Seber (1996) et illustrés dans Chao et Thompson (2001).

Bibliographie

- Dedrick, C.L. (1938). *Census of unemployment 1937: Principle findings of the enumerative check census*. U.S. Bureau of the Census.
- Duncan, J.W., et Shelton, W.C. (1978). *Revolution in United States Government Statistics 1926 - 1976*. U.S. Department of Commerce.
- Frankel, L.R., et Stock, J.S. (1942). On the sample survey of unemployment. *Journal of the American Statistical Association*, 37, 77-80.

- Stephan, F.F., Deming, W.E. et Hansen, M.H. (1940). The sampling procedure of the 1940 population census. *Journal of the American Statistical Association*, 35, 615-630.
- Stephan, F.F. (1948). History of the uses of modern sampling procedures. *Journal of the American Statistical Association*, 43, 12-39.
- U.S. Bureau of Labor Statistics et U.S. Census Bureau (2006). *Design and Methodology: Current Population Survey*.

« Autres plans de sondage : Échantillonnage avec bases de sondage multiples chevauchantes », par Sharon Lohr ; « Dix années d'échantillonnage équilibré par la méthode du cube : une évaluation », par Yves Tillé ; « Plans d'échantillonnage novateurs : discussion de trois communications présentées au U.S. Census Bureau », par Jean Opsomer.

4. Prochaines étapes

Après les trois exposés, il a été décidé de poursuivre l'étude de ces méthodes et de leur application aux enquêtes démographiques existantes du U.S. Census Bureau ou à de nouvelles enquêtes éventuelles. Il existe déjà un besoin urgent d'appliquer les méthodes à bases de sondage chevauchantes multiples à la National Survey of College Graduates en vue de résoudre un problème d'ancienne cohorte/nouvelle cohorte et une utilisation éventuelle des registres des permis de chasse et de pêche des États comme deuxième base de sondage pour la Fishing, Hunting, and Wildlife-Associated Recreation survey. Nous prévoyons examiner l'échantillonnage équilibré, surtout pour sélectionner les unités primaires d'échantillonnage géographiques. Enfin, les méthodes d'échantillonnage adaptatif pourraient nous permettre d'accepter des enquêtes que nous n'entre-

prenons pas habituellement, et pourraient représenter une option moins onéreuse pour les enquêtes qui satisfont certains critères.

5. Résumé

La présente exploration de ces trois catégories de plans de sondage de rechange ne constitue que le début de notre série de séminaires et de notre examen des moyens d'améliorer nos méthodes d'élaboration du plan de sondage des enquêtes démographiques. Les sujets que nous prévoyons aborder dans l'avenir comprennent d'autres méthodes de production de listes, l'approche des intervalles demi ouverts de Kish pour les mises à jour de la croissance et l'amélioration de la couverture, les plans d'enquête adaptatifs, les méthodes d'échantillonnage réjectif et l'échantillonnage assisté d'un modèle.

Remerciements

Le présent rapport est diffusé en vue de tenir les parties intéressées au courant des travaux de recherche et de favoriser les discussions. Les opinions exprimées en ce qui concerne les questions statistiques, méthodologiques, techniques ou opérationnelles sont celles des auteurs et ne reflètent pas forcément celles du U.S. Census Bureau.

Schools and Staffing Survey, de la Private School Survey et de la Survey of Inmates of Local Jails. D'autres encore sont réalisées auprès d'échantillons stratifiés tirés d'une base de sondage, comme la National Survey of College Graduates, dont l'échantillon a été sélectionné parmi les répondants au questionnaire complet du recensement décennal, et l'American Time Use Survey, dont les échantillons sont tirés de la CPS. Au début des années 1990, le U.S. Census Bureau a entrepris des travaux de développement en vue d'utiliser une mesure continue pour remplacer éventuellement la version longue du questionnaire du recensement décennal. Depuis, ces efforts ont abouti à l'American Community Survey courante, qui, à partir de 2010, fournira des estimations mi-décennales continues jusqu'au niveau du groupe d'îlots. L'objectif du Census Bureau est d'améliorer ses méthodes d'échantillonnage, ce qui nous amène maintenant à l'étude d'autres plans de sondage.

3. Série de séminaires sur d'autres plans de sondage

L'exploration d'autres méthodes d'échantillonnage a débuté par une première série de séminaires qui s'est déroulée au U.S. Census Bureau. Elle comprenait trois séminaires consacrés à ce genre de méthodes durant lesquels ont été examinés les fondements statistiques des méthodes et leurs limites, surtout dans le contexte de leur application aux types d'enquêtes démographiques menées par le U.S. Census Bureau. Chaque exposé a aussi été commenté par le professeur Jean Opsomer de la Colorado State University. Par la suite, trois articles ont été rédigés en vue de fournir des renseignements plus détaillés sur chaque sujet, ainsi qu'un article final du commentateur portant sur les trois sujets abordés durant les séminaires.

- Le 26 septembre 2007, le professeur Steven K. Thompson de la Simon Fraser University a présenté un exposé sur ses travaux dans le domaine de l'échantillonnage en réseau, de l'échantillonnage spatial et de l'échantillonnage adaptatif.
- Le 9 janvier 2008, la professeure Sharon Lohr de l'Arizona State University a présenté un exposé sur ses travaux dans le domaine de l'échantillonnage au moyen de bases de sondage chevauchantes.
- Le 4 juin 2008, le professeur Yves Tillé de l'Université de Neuchâtel a présenté un exposé sur ses travaux de recherche dans le domaine de l'échantillonnage équilibré.

Les articles émanant de ce projet sont les suivants :

« Échantillonnage adaptatif par réseau et spatial », par Steven Thompson ;

Autres plans de sondage pour les enquêtes démographiques étudiées par le U.S. Census Bureau

Patrick E. Flanagan et Ruth Ann Killian¹

1. Introduction

Le Programme de remaniement des échantillons des enquêtes démographiques du U.S. Census Bureau est chargé, entre autres, de mener des travaux de recherche afin d'améliorer la conception des enquêtes démographiques menées aux États-Unis, en se penchant particulièrement sur les plans de sondage. Historiquement, la recherche en vue d'améliorer les plans de sondage s'est limitée aux méthodes « classiques » telles que la stratification élémentaire, l'échantillonnage à plusieurs degrés, l'échantillonnage systématique, l'échantillonnage avec probabilité proportionnelle à la taille, l'échantillonnage en grappes et l'échantillonnage aléatoire simple. Au cours des quelque 30 dernières années, nous n'avons cessé de voir baisser les taux de réponse et grimper les coûts, tandis qu'augmentait la demande de données sur tous les types de population. Plus récemment, l'accroissement spectaculaire de la puissance informatique et l'accès à des données connexes provenant de dossiers administratifs ont laissé entendre que nous pourrions disposer de plus d'options aujourd'hui que nous n'en avions au moment où nous avons établi notre méthodologie courante. Nous avons donc entrepris un projet visant à étudier d'autres méthodes d'échantillonnage.

2. Historique de l'innovation dans le domaine de l'échantillonnage démographique au U.S. Census Bureau

Le U.S. Census Bureau a été créé aux termes de la *Permanent Census Act* de 1902. Jusqu'à la fin des années 1930, les travaux démographiques du Bureau concernaient principalement la logistique de l'exécution de chaque recensement décennal et d'une myriade de recensements spéciaux. Après le recensement décennal de 1930, le Census Bureau a entrepris des travaux de recherche sur l'échantillonnage en se servant des données de recensement (Stephan 1948).

Puis, en 1937, le Census Bureau a fait son premier pas important dans le domaine de l'échantillonnage avec l'Enumerative Check Census of Unemployment de 1937, dans le cadre duquel un échantillon en grappes de comtés a été utilisé pour appuyer un recensement des registres du chômage (Dedrick 1938). À peu près au même moment,

l'extension du recensement décennal, des spécialistes de l'échantillonnage ont été recrutés (dont W. Edwards Deming et Frederick Stephan) afin qu'ils participent à la conception d'une enquête par sondage qui serait conjuguée au recensement décennal de 1940 et fondée sur un échantillon systématique au 1/20 (Stephan, Deming et Hansen 1940). En 1942, la Sample Survey of Unemployment a été transférée de la Works Progress Administration au U.S. Census Bureau. Cette enquête était déjà menée auprès d'un échantillon à trois degrés comprenant la sélection des comtés comme unités primaires d'échantillonnage (UPE). L'échantillonnage systématique d'îlots et l'échantillonnage d'après les listes d'unités de logement au troisième degré (Frankel et Stock 1942). Après son transfert au Census Bureau (et un changement de nom pour devenir le Monthly Report on the Labor Force ou MRLF), l'enquête a été remaniée en profondeur en 1943 et son efficacité a été considérablement améliorée grâce à des unités primaires d'échantillonnage (UPE) plus grandes et des probabilités de sélection proportionnelles à la taille (Duncan et Shelton 1978). Plus tard, l'enquête a été modifiée afin d'améliorer les comparaisons d'un mois à l'autre et d'une année à l'autre en utilisant une approche plus complexe d'échantillons chevauchants dans lesquels un ménage particulier restait dans l'échantillon pendant quatre mois, en sortait pendant huit mois, puis y entrait de nouveau pendant quatre mois. En 1947, elle a été renommée pour devenir la Current Population Survey (CPS). Néanmoins, le concept d'échantillonnage de base demeurerait l'échantillonnage à plusieurs degrés dans lequel l'UPE était le comté ou le groupe de comtés. Il en est encore ainsi aujourd'hui, quoiqu'il existe des différences majeures en ce qui a trait aux méthodes d'échantillonnage à l'intérieur des UPE (U.S. Bureau of Labor Statistics et U.S. Census Bureau 2006). Au cours des 60 dernières années, le U.S. Census Bureau a conçu de nombreuses enquêtes démographiques supplémentaires. Certaines de ces enquêtes reposent sur la même notion d'échantillonnage à deux degrés que celle utilisée dans la CPS, comme les Consumer Expenditures Surveys, la National Survey of Income and Program Participation, la National Crime Victimization Survey et la National Health Interview Survey. D'autres comprennent un plan de sondage à deux degrés avec sélection de listes d'origine, suivies d'un échantillonnage à partir des listes, comme dans le cas de la

SEVANI (Beaumont, Bissonnette et Bocci 2010), que l'on peut obtenir sur demande auprès des auteurs.

Remerciements

Nous tenons à remercier les réviseurs de leurs commentaires. Nous remercions également Mike Hidiroglou, Eric Rancourt et Cynthia Bocci de Statistique Canada de leurs suggestions, sans oublier les discussions tenues sur le sujet. Tous ces commentaires ont servi à peaufiner notre article.

Bibliographie

Beaumont, J.-F., Bissonnette, J. et Bocci, C. (2010). SEVANI, version 2.3, Guide de méthodologie. Rapport interne, Direction de la méthodologie. Statistique Canada.

Beaumont, J.-F., et Bocci, C. (2009). Variance estimation when donor imputation is used to fill in missing values. *Canadian Journal of Statistics*, 37, 400-416.

Beaumont, J.-F., Haziza, D. et Bocci, C. (2011). On variance estimation under auxiliary value imputation in sample surveys. *Statistica Sinica*, 21, 515-537.

Brick, J.M., Kalton, G. et Kim, J.K. (2004). Estimation de variance pour l'imputation hot deck à l'aide d'un modèle. *Techniques d'enquête*, 30, 63-72.

Deville, J.-C., et Särndal, C.-E. (1994). Variance estimation for the regression imputed Horvitz-Thompson estimator. *Journal of Official Statistics*, 10, 381-394.

Felix, P., et Rancourt, E. (2001). Applications de la variance due à l'imputation dans l'Enquête sur l'emploi, la rémunération et les heures. Document de travail, Direction de la méthodologie, Statistique Canada, BSMID-2001-009E.

Haziza, D. (2009). Imputation and inference in the presence of missing data. Dans *Handbook of Statistics, Sample Surveys: Theory, Methods and Inference*, (Eds., D. Pfeffermann et C.R. Rao). Amsterdam : Elsevier BV, 29A, 215-246.

Hidiroglou, M.A. (1989). Notes manuscrites non publiées générées par l'auteur.

Kim, J.-K., et Rao, J.N.K. (2009). Unified approach to linearization variance estimation from survey data after imputation for item nonresponse. *Biometrika*, 96, 917-932.

Lee, H., Rancourt, E. et Särndal, C.-E. (2001). Variance estimation from survey data under single imputation. Dans *Survey Nonresponse*, (Eds., R.M. Groves, D.A. Dillman, J.L. Eltinge et R.J.A. Little). New-York : John Wiley & Sons, Inc., 315-328.

Rancourt, E., Lee, H. et Särndal, C.-E. (1993). Variance estimation under more than one imputation method. Dans *Proceedings of the International Conference on Establisshments Surveys*, juin 1993, Buffalo, American Statistical Association, 374-379.

Rubin, D.B. (1987). *Multiple Imputation for Nonresponse in Surveys*. New-York : John Wiley & Sons, Inc.

Särndal, C.-E. (1992). Méthodes pour estimer la précision des estimations d'une enquête ayant fait l'objet d'une imputation. *Techniques d'enquête*, 18, 257-268.

Shao, J., et Steel, P. (1999). Variance estimation for survey data with composite imputation and nonnegligible sampling fractions. *Journal of the American Statistical Association*, 94, 254-265.

Sitter, R.R., et Rao, J.N.K. (1997). Imputation for missing values and corresponding variance estimation. *Canadian Journal of Statistics*, 25, 61-73.

donné que $V_{MC}^{SAM} / V_{MC}^{TOT}$ n'est pas proche de 100 %, même pour $n = 100$, les effets de la non-réponse et de l'imputation ne peuvent être systématiquement omis lorsque l'on estime l'EQM globale.

7. L'approche inverse

Shao et Steel (1999) ont proposé une approche inverse d'estimation de la variance, élaborée pour les situations où l'on a recours à l'imputation composite. Ils ont fait l'hypothèse que le biais global est négligeable et ont mis de l'avant la décomposition suivante de la variance globale :

$$E_{mpq}(\theta_I - \theta)^2 = E_{mpq} \text{var}_p(\theta_I | U_p) + E_{mq}\{E_p(\theta_I | U_p) - \theta\}^2, \quad (7.1)$$

où U_p est une population conceptuelle de répondants. Dans le membre droit de l'expression (7.1), l'espérance et la variance intérieures sont déterminées par rapport au plan d'échantillonnage. Malheureusement, l'estimateur imputé θ_I ne sera généralement pas linéaire par rapport à ce plan, même s'il l'est par rapport aux valeurs de y observées. Par conséquent, l'estimateur imputé θ_I est généralement linéarisé (se reporter par exemple à Shao et Steel 1999, ainsi qu'à Kim et Rao 2009). De manière plus explicite, les quantités $\phi_{(j)}^{0k}$ et $\phi_{(j)}^{lk}$ dépendent souvent de l'échantillon, et ce, de façon non linéaire ; c'est notamment le cas pour l'imputation par la régression linéaire (se reporter à l'exemple présenté à la fin de la section 3) et pour l'imputation par donneur. Il n'est pas toujours facile de prendre en compte la variabilité de $\phi_{(j)}^{0k}$ et de $\phi_{(j)}^{lk}$ lorsque l'on utilise (7.1). Par exemple, il n'y a pas d'articles portant sur l'utilisation de l'approche inverse pour estimer la variance lorsque l'on a recours à l'imputation par le plus proche donneur. De plus, étant donné que chaque stratégie d'imputation composite produit son propre estimateur imputé linéarisé, il est ardu de mettre en oeuvre cette méthodologie dans un progiciel généralisé.

est calculée par rapport au modèle d'imputation (conditionnellement à s et s_p). L'estimateur imputé est linéaire et les calculs sont simples parce que les quantités $\phi_{(j)}^{0k}$ et $\phi_{(j)}^{lk}$ sont établies sans faire intervenir les valeurs de y . Ces deux quantités n'entrent pas dans l'estimation de la variance d'échantillonnage, $V_{SAM}^{var} = E_m \text{var}_p(\theta)$ (se reporter à l'équation 5.5), de sorte que leur éventuelle non-linéarité par

et à la composante mixte,

$$V_{MX}^{var} = 2E_{pq}E_m\{(\theta_I - \theta)(\hat{\theta}_I - \theta) | s, s_p\},$$

$$V_{NR}^{var} = E_{pq}E_m\{(\theta_I - \theta)^2 | s, s_p\},$$

les expressions relatives à la variance due à la non-réponse,

Au moyen de notre approche, l'espérance intérieure dans

8. Conclusion

Notre méthodologie relative à l'imputation composite a été mise en oeuvre dans la version 2 du SEVANI en raison de sa facilité d'exécution et de sa généralité. Elle fonctionne pour la plupart des méthodes d'imputation utilisées dans la pratique, car la grande majorité de ces méthodes sont linéaires. Les calculs de variance sont les mêmes pour toutes les stratégies d'imputation composite, une fois que l'on a calculé les quantités $W_{0(+)}^{0d}$, $W_{0(+)}^{dk}$, μ_k et σ_k^2 . Il est ainsi plus facile de procéder au développement d'un système généralisé.

Nous avons mis l'accent sur l'estimation d'un total de domaine au moyen de l'estimateur de Horvitz-Thompson, mais le SEVANI peut aussi être utilisé relativement à des estimateurs de moyennes de domaine et à des estimateurs par calage. Il y a aussi des méthodes paramétriques et non paramétriques d'estimation de μ_k et σ_k^2 . On trouvera de plus amples détails dans le guide méthodologique du

rapport au plan d'échantillonnage ne pose aucun problème. Cela veut dire que l'imputation par le plus proche donneur peut être traitée facilement avec notre approche (se reporter à Beaumont et Bocci 2009).

Ce sont toutes ces raisons qui nous amènent à penser que l'approche inverse risque d'être plus fastidieuse à mettre en oeuvre dans un progiciel généralisé que notre approche. Cela ne veut pas dire que l'approche inverse est inutile. Dans les faits, les deux approches aboutissent à des estimateurs de variance identiques lorsqu'un recensement est effectué. Beaumont, Haziza et Bocci (2011) ont montré que l'une et l'autre approches donnent également des estimateurs de variance identiques lorsque l'on utilise l'imputation par valeur auxiliaire (étant donné que $\phi_{(j)}^{0k}$ et $\phi_{(j)}^{lk}$ ne dépendent pas de s et s_p). Les deux approches dépendent d'une spécification correcte du modèle d'imputation, et aucune des deux ne devrait donner systématiquement de meilleurs résultats que l'autre.

L'approche inverse peut avoir un avantage pratique comparativement à la nôtre quand la fraction de sondage est négligeable. Dans un tel cas, Shao et Steel (1999) montrent que la deuxième composante du membre droit de (7.1) peut être négligée. Pour estimer la première composante, on détermine un estimateur fondé sur le plan de sondage pour $\text{var}_p(\theta_I | U_p)$. Si l'on choisit une méthode d'estimation de la variance par réplication (comme le jackknife ou le bootstrap) pour estimer $\text{var}_p(\theta_I | U_p)$, l'approche dans son ensemble devient fort intéressante et pratique. En outre, elle ne dépend pas de la validité du modèle d'imputation, en particulier la spécification correcte de la variance sous le modèle σ_k^2 . Les estimateurs de la variance par le jackknife de Rancourt, Lee et Särndal (1993) et de Sitter et Rao (1997) peuvent être justifiés par cette approche.

titre d'exemple, nos conditions de simulation sont pour l'essentiel les mêmes que celles commentées par Rancourt, Lee et Särndal (1993), qui se sont aussi penchés sur l'imputation composite.

Nous avons calculé la variance d'échantillonnage Monte Carlo et l'EQM globale, $V_{MC}^{SAM} = \sum_{r=1}^R (\theta_r - \theta)^2 / R$, respectivement, où l'indice r indique que les estimations sont fondées sur la r^e répétition, et $R = 10\,000$. Le biais relatif Monte Carlo de tout estimateur de V_{SAM}^{SAM} , par exemple $V_{SAM}^{MC} / (V_{MC}^{SAM} / V_{TOT}^{SAM})$, est calculé ainsi : $RB(V_{SAM}^{TOT}) = \sum_{r=1}^R (V_{SAM,r}^{TOT} - V_{MC}^{SAM}) / (V_{MC}^{SAM} / V_{TOT}^{SAM})$. De même, nous avons calculé le biais relatif Monte Carlo d'un estimateur de V_{TOT}^{TOT} , désigné comme étant $RB(V_{TOT}^{TOT})$, et celui d'un estimateur de $V_{TOT}^{TOT} / V_{TOT}^{TOT}$, soit $RB(V_{TOT}^{TOT} / V_{TOT}^{TOT})$. Enfin, nous avons calculé les taux de couverture de Monte Carlo des intervalles de confiance à 95 % pour θ , en supposant une distribution normale de θ_I .

Les résultats de notre étude par simulation sont présentés au tableau 2. Dans les colonnes identifiées SEVANI, la variance d'échantillonnage, V_{SAM}^{SAM} , et l'EQM globale, V_{TOT}^{TOT} , sont estimées pour chaque échantillon au moyen de V_{TOT}^{TOT} et V_{TOT}^{TOT} , respectivement (se reporter à la section 5.4). Nous avons également obtenu des résultats en remplaçant V_{TOT}^{TOT} par V_{TOT}^{TOT} . Nous ne les présentons toutefois pas dans le tableau 2, car ils étaient très près de ceux obtenus avec V_{TOT}^{TOT} . Cela donne à penser que le biais par rapport au modèle, B_m , n'est pas important ici. Dans les colonnes Naïf, tant la variance d'échantillonnage que l'EQM globale sont estimées au moyen de V_{ORD}^{ORD} (se reporter à la section 5.1).

Tableau 2
Résultats de l'étude par simulation

$n = 100$			$n = 250$		
SEVANI	Naïf	SEVANI	Naïf	SEVANI	Naïf
$RB(V_{SAM}^{TOT})$	2,82 %	-17,59 %	3,02 %	-17,68 %	-
$RB(V_{SAM}^{TOT} / V_{TOT}^{TOT})$	8,30 %	-	5,84 %	-52,89 %	-
Taux de couverture	93,38 %	86,20 %	94,42 %	81,80 %	

Ces résultats montrent que la méthodologie décrite à la section 5 et exécutée dans le SEVANI est plus efficace que l'estimateur naïf de la variance et d'établir des intervalles de confiance. L'utilisation du SEVANI donne de petits biais relatifs Monte Carlo et des taux de couverture proches du taux nominal cible (95 %). Notre méthodologie est aussi utile aux utilisateurs voulant estimer la part de l'EQM globale attribuable à la variance d'échantillonnage, c'est-à-dire $V_{SAM}^{SAM} / V_{TOT}^{TOT}$. Il est à noter que $V_{MC}^{SAM} / V_{TOT}^{TOT}$ est de 71,98 % pour $n = 100$ et de 57,23 % pour $n = 250$. Étant

un résidu aléatoire a été ajouté aux fins d'imputation de l'unité k , à défaut de quoi $r_k = 0$. L'estimateur imputé (2.1), dans lequel y_k^* est remplacé par $y_k^* R_k$, est désigné par $\theta_I^* = \theta_I + \sum_{k \in s_m} w_k d_k r_k e_k$. Étant donné que $E_k(e_k | s, s_r) = 0$, l'ajout d'un résidu aléatoire n'entraîne aucun biais dans l'estimateur imputé. L'EQM globale de θ_I^* peut être exprimée ainsi :

$$E_{mpq}^{mpq}(\theta_I^* - \theta)^2 = E_{mpq}(\theta_I - \theta)^2 + E_{mpq} \text{var}_s(\theta_I^* | s, s_r). \quad (5.13)$$

Nous estimons le premier terme du membre droit de (5.13) comme à la section 5.4. Nous estimons le deuxième terme par :

$$\text{var}_s(\theta_I^* | s, s_r) = \sum_{k \in s_m} w_k^2 d_k r_k \hat{\sigma}_k^2. \quad (5.14)$$

6. Étude par simulation

Nous avons effectué une étude par simulation Monte Carlo pour évaluer la méthodologie décrite à la section 5. Une population bivariable de $N = 400$ unités a été générée contenant une variable auxiliaire x et une variable d'intérêt y . Pour chaque unité que compte la population, la variable auxiliaire a été générée selon une loi gamma de moyenne 48 et de variance 768. La variable d'intérêt a été générée conditionnellement à x selon une loi gamma de moyenne 1,5 x et de variance 16 x . On a attribué de façon aléatoire une valeur manquante de x à la moitié de la population. Étant donné qu'aucun domaine d'intérêt n'a été généré, θ correspond au total de population de la variable y .

Nous avons sélectionné 10 000 échantillons à partir de cette population par échantillonnage aléatoire simple sans remise. Nous avons considéré deux tailles d'échantillon : $n = 100$, et $n = 250$. Pour chaque échantillon, la non-réponse associée à la variable y a été générée indépendamment d'une unité à l'autre, la probabilité de non-réponse étant de 0,3. Nous avons utilisé la même stratégie d'imputation que dans l'exemple de la section 2, avec $\omega^{(1)} = 1$, pour $l \in s_r^{(1)}$, et $\omega^{(2)} = 1$, pour $l \in s_r$. Les non-répondants pour la variable y avec valeur de x observée ont été imputés au moyen de la méthode d'imputation par le ratio, l'imputation par la moyenne étant utilisée lorsque la valeur de x était manquante.

Les valeurs de y de la population sont demeures fixes tout au long des répétitions de l'expérience par simulation ; chaque répétition consistait à sélectionner un échantillon, puis à générer la non-réponse à l'égard de la variable y . Si nous nous en étions tenus strictement au développement théorique exposé à la section 5, nous aurions généré de nouvelles valeurs de y lors de chaque répétition conformément au modèle d'imputation. Il est toutefois plus courant dans la littérature de fixer les valeurs de y de la population lorsque l'on procède à une expérience par simulation. À

Pour obtenir l'estimateur B_m , nous remplaçons μ_k dans la formule (5.10) par un estimateur convergent sous m, μ_k .

5.3 Estimation de la composante mixte

On obtient un estimateur sans biais par rapport à m, p et q de la composante mixte

$$V_{\text{MIX}} = 2E_{pq} E_m \{ (\theta_I - \theta) (\theta - \theta) | s, s_p \}$$

en trouvant un estimateur sans biais par rapport à m de :

$$2E_m \{ (\theta_I - \theta) (\theta - \theta) | s, s_p \} = 2\text{cov}_m \{ (\theta_I - \theta), (\theta - \theta) | s, s_p \} + 2B_m E_m \{ (\theta - \theta) | s, s_p \}. \quad (5.11)$$

Puisque l'erreur due à la non-réponse et l'erreur d'échantillonnage sont linéaires dans les valeurs de y , l'utilisation de (5.8) donne :

$$2\text{cov}_m \{ (\theta_I - \theta) (\theta - \theta) | s, s_p \} = 2 \sum_{k \in s_p} W_{(+)k}^{\text{dk}} (w_k - 1) d_k \sigma_k^2 - 2 \sum_{k \in s_p} w_k (w_k - 1) d_k \sigma_k^2. \quad (5.12)$$

Si le biais par rapport au modèle, B_m , est négligeable, on obtient un estimateur V_{MIX} sans biais par rapport à m, p , et q de la composante mixte, V_{MIX} , en remplaçant σ_k^2 dans l'équation (5.12) par un estimateur sans biais par rapport à m (et convergent sous m), $\hat{\sigma}_k^2$. Il est à noter que la composante mixte ne sera pas forcément négligeable (Brick, Kalton et Kim 2004) et qu'elle est même souvent négative dans la pratique.

Si le biais par rapport au modèle, B_m , n'est pas négligeable, il ne sera peut-être pas possible d'estimer facilement la deuxième composante du membre droit de (5.11), parce que $E_m \{ (\theta - \theta) | s, s_p \}$ requiert que l'on connaisse x_{obs}^k ainsi que la variable indicatrice de domaine d pour la partie non échantillonnée de la population, or cette information pourrait ne pas être disponible. Il est possible de surmonter le problème en modifiant le cadre inferentiel. Nous pouvons modéliser la distribution multivariée complète reliant y, x et d , au lieu de conditionner sur d et x_{obs} . Nous n'avons pas joué cette idée dans le SEVANI parce qu'elle entraîne une tâche de modélisation plus complexe et qu'il est difficile d'obtenir une expression générale de la variance facile à mettre en œuvre. En pratique, si le biais par rapport au modèle n'est pas trop grand, le fait d'ignorer la deuxième composante du membre droit de (5.11) ne devrait pas causer trop de souci. À la section 5.4, nous proposons une statistique qui peut aider à déterminer si le biais par rapport au modèle est important ou non.

La composante mixte peut aussi être exprimée sous la forme suivante :

$$V_{\text{MIX}} = 2E_{pq} E_m \{ (\theta_I - \theta) (\theta - \theta) | s, s_p \} + 2E_{pq} [\text{cov}_m \{ (\theta_I - \theta), (\theta - \theta) | s, s_p \}] + 2E_{pq} [E^q(B_m | s) E_m \{ (\theta - \theta) | s \}].$$

L'expression (5.12) peut des lors être utilisée pour obtenir un estimateur de V_{MIX} , à condition que $E^q(B_m | s)$ soit négligeable. Cette hypothèse est plus faible que le fait d'exiger que B_m soit négligeable, puisqu'on y satisfait si B_m ou $E^q(\theta_I - \theta | s)$ est négligeable. Par exemple, dans notre exemple précédent, B_m peut ne pas être négligeable mais, si $d_k = 1$ et $w_k^{(2)} = w_k$, $E^q(\theta_I - \theta | s) \approx 0$ sous une non-réponse uniforme (se reporter à Sitter et Rao 1997).

5.4 Estimation de l'EQM globale/de la variance globale

L'EQM globale, ou la variance globale si le biais global est négligeable,

$$V^{\text{TOT}} = E_{mpq} (\theta_I - \theta)^2 = V^{\text{SAM}} + V^{\text{NR}} + V^{\text{MIX}}$$

peut être estimée au moyen de $V^{\text{TOT}} = V^{\text{SAM}} + V^{\text{NR}} + V^{\text{MIX}}$ si le biais par rapport au modèle, B_m , est négligeable. L'estimateur de la composante de la non-réponse est $V^{\text{NR}} + V^{\text{MIX}}$. Du point de vue de l'utilisateur, l'estimateur V^{TOT} présente plus d'intérêt que ses composantes. Cela étant, l'utilisateur pourrait s'intéresser à l'estimateur de la variance d'échantillonnage, V^{SAM} , ou au ratio $V^{\text{SAM}} / V^{\text{TOT}}$. Ce dernier sert à estimer l'apport de la variance d'échantillonnage à la variance globale. Tel qu'indiqué à la section 5.2, si le biais par rapport au modèle n'est pas négligeable, la variance de la non-réponse peut être estimée par $V^{\text{NR}} + B_m^2$ plutôt que par V^{NR} . L'estimateur de l'EQM globale est alors $V^{\text{TOT, ADJ}} = V^{\text{SAM}} + (V^{\text{NR}} + B_m^2) + V^{\text{MIX}}$.

Une statistique pouvant être utile à titre de diagnostic pour déterminer l'ampleur du biais sous le modèle est $|B_m| / \sqrt{V^{\text{TOT}}}$, ou encore $|B_m| / \sqrt{V^{\text{TOT, ADJ}}}$. Si l'une ou l'autre de ces statistiques présente une valeur élevée, cela peut indiquer que le biais sous le modèle n'est pas négligeable et que la procédure d'imputation composite doit être remise en question. Comparativement à $|B_m| / \sqrt{V^{\text{TOT}}}$, $|B_m| / \sqrt{V^{\text{TOT, ADJ}}}$ présente l'avantage d'être bornée, c'est-à-dire :

$$0 \leq |B_m| / \sqrt{V^{\text{TOT, ADJ}}} \leq 1.$$

5.5 Imputation par la régression aléatoire

Un résidu de régression aléatoire e_k est parfois ajouté à la valeur imputée par régression y_k^* afin de préserver la variabilité naturelle de la variable y . Nous proposons de générer indépendamment les résidus aléatoires e_k avec $E(e_k | s, s_p) = 0$ et $\text{var}_k(e_k | s, s_p) = \hat{\sigma}_k^2$, où l'indice * indique que l'espérance et la variance sont considérées par rapport au mécanisme d'imputation aléatoire. Nous obtenons ainsi la valeur imputée $y_k^* = y_k^* + r_k e_k$, où $r_k = 1$ si

On peut montrer que, dans des conditions faibles, $E_m^*(\beta_2^* | s, s_r) = \beta_2 + O_p(1/\sqrt{n})$, de sorte que le biais de β_2^* sous le modèle est asymptotiquement négligeable. Toutefois, étant donné que $\text{var}_m(\beta_2^* | s, s_r) = O_p(1/n)$, le biais quadratique du modèle n'est pas nécessairement asymptotiquement négligeable par rapport à la variance de β_2^* sous le modèle. Au moins, β_2^* est convergent sous m pour β_2 . On peut voir à partir de l'équation (4.3) ou (4.4) qu'il est possible de contrôler le biais de β_2^* sous le modèle en appliquant un poids $\omega_k^{(2)}$ plus faible aux unités $k \in s_r^{(1)}$ qu'aux unités $k \in s_r^{(2)}$. Par exemple, on pourrait envisager d'utiliser $\omega_k^{(2)} = w_k/n^\alpha$, pour $k \in s_r^{(1)}$ et une certaine valeur de $\alpha > 0$, et $\omega_k^{(2)} = w_k$, pour $k \in s_r^{(2)}$. Dans le cas extrême où $\omega_k^{(2)} = 0$, pour $k \in s_r^{(1)}$, β_2^* est sans biais sous le modèle, car il est égal à β_2 . Précisons que le biais de β_2^* sous le modèle pourrait être supérieur à $O_p(1/\sqrt{n})$ si x_{1k} , $k \in s_r^{(1)}$, présentent une moyenne différente de x_{1k} , $k \in s_r^{(2)}$. Il pourrait alors être plus important de contrôler le biais de β_2^* sous le modèle.

Dans le cas de l'imputation par donneur, il faut tenir compte d'une quatrième source de variabilité quand les donneurs sont sélectionnés aléatoirement parmi les répondants pour imputer les données des non-répondants. Dans cet article, l'indice q indiquera implicitement que les moments sont évalués par rapport à la distribution conjointe induite par le mécanisme de non-réponse et le mécanisme de sélection aléatoire des donneurs. Par conséquent, lorsque nous procédons au conditionnement sur s_r , comme dans l'équation (4.2), il faut se rappeler que le conditionnement est effectué non seulement sur l'ensemble de répondants, mais aussi sur l'ensemble de donneurs sélectionnés.

5. Estimation de la variance

Sämdal (1992) exprime l'erreur totale de l'estimateur imputé sous la forme suivante :

$$\theta_j - \theta = (\hat{\theta} - \theta) + (\hat{\theta}_j - \theta), \quad (5.1)$$

où le premier terme du membre droit de (5.1) est appelé erreur d'échantillonnage et le deuxième, erreur due à la non-réponse. Si nous reprenons les hypothèses de la section 4, et à partir du moment où $E_p(\hat{\theta} - \theta) = 0$, le biais global de l'estimateur imputé se réduit à $E_{mpq}(\hat{\theta}_j - \theta) = E_{pq}B_m$, où $B_m = E_m(\hat{\theta}_j - \theta | s, s_r)$ est le biais (conditionnel) de l'estimateur imputé par rapport au modèle. À partir de (2.1), le biais par rapport au modèle peut s'exprimer sous la forme suivante :

$$B_m = \sum_j \sum_{k \in s_r^{(j)}} w_k d_k E_m(y_k^* - y_k | s, s_r). \quad (5.2)$$

Cela signifie que le biais par rapport au modèle et le biais global disparaissent si l'espérance sous le modèle de l'erreur d'imputation, $y_k^* - y_k$, est nulle, pour $k \in s_r^{(j)}$ et $j = 1, \dots, J$. En principe, la stratégie d'imputation doit être choisie de sorte que cette condition soit remplie, au moins approximativement. C'est une hypothèse courante dans la littérature (par exemple, Sämdal 1992 ; Shao et Steel 1999). Dans l'exemple amorcé à la section 2, le biais par rapport au modèle (5.2) se réduit à :

$$B_m = \left(\sum_{k \in s_r^{(2)}} w_k d_k \right) E_m(\beta_2^* - \beta_2 | s, s_r).$$

L'expression de $E_m(\beta_2^* - \beta_2 | s, s_r)$ peut être donnée par (4.4) ou (4.4). Tel que mentionné dans le paragraphe qui suit l'équation (4.4), on peut contrôler le biais par rapport au modèle, B_m , en appliquant un poids $\omega_k^{(2)}$ plus faible aux unités $k \in s_r^{(1)}$ qu'aux unités $k \in s_r^{(2)}$. Le biais sera également peu marqué si le nombre de non-répondants à l'égard desquels on a procédé à l'imputation au moyen de la méthode 2 est peu élevé. Il est à noter que notre approche d'estimation de la variance (ou de l'erreur quadratique moyenne, ou EQM) requiert l'hypothèse un peu plus faible selon laquelle $E_q(B_m | s)$ est négligeable (se reporter à la section 5.3).

À partir de (5.1), Sämdal (1992) décompose l'EQM globale en trois composantes :

$$E_{mpq}(\hat{\theta}_j - \theta)^2 = E_m \text{var}_p(\hat{\theta}) + E_{pq} E_m \{(\hat{\theta}_j - \theta)^2 | s, s_r\} + 2E_{pq} E_m \{(\hat{\theta}_j - \theta)(\hat{\theta} - \theta) | s, s_r\}. \quad (5.3)$$

L'EQM globale (5.3) équivaut dès lors à peu près à la variance globale, $\text{var}_{mpq}(\hat{\theta}_j - \theta)$, lorsque le biais global est négligeable. Les premier, deuxième et troisième termes du membre droit de (5.3) sont appelés variance d'échantillonnage, variance due à la non-réponse et composante mixte, respectivement. La somme des deux derniers termes peut être appelée composante due à la non-réponse, car ces termes disparaissent en l'absence de non-réponse. La composante due à la non-réponse correspond simplement à la différence entre l'EQM – ou la variance – globale et la variance d'échantillonnage. Nous allons ci-après élaborer un estimateur pour chacun de ces trois termes.

5.1 Estimation de la variance d'échantillonnage

Soit $v(y)$, un estimateur de $\text{var}_p(\hat{\theta})$ sans biais par rapport à p que nous utilisons s'il y avait réponse complète. L'estimateur de Horvitz-Thompson typique est :

$$v(y) = \sum_{k \in s} \sum_{kl} \frac{\pi_{kl}}{\pi_{kl} - \pi_k \pi_l} (w_k d_k y_k)(w_l d_l y_l). \quad (5.4)$$

suite de l'article, nous utiliserons les indices m , p et q pour désigner les espérances, les variances et les covariances évaluées par rapport au modèle d'imputation, au plan d'échantillonnage et au mécanisme de non-réponse, respectivement. Nous prenons le modèle d'imputation suivant pour décrire la relation entre la variable y et le vecteur de variables auxiliaires observées \mathbf{x}^{obs} :

$$(4.1) \quad \begin{aligned} E_m(y_k | \mathbf{x}^{\text{obs}}) &= \mu_k \\ V_m(y_k | \mathbf{x}^{\text{obs}}) &= \sigma_k^2 \\ \text{cov}_m(y_k, y_l | \mathbf{x}^{\text{obs}}) &= 0, \end{aligned}$$

pour $k \neq l$ et $k, l \in U$. La matrice de population, \mathbf{X}^{obs} , contient les vecteurs de variables auxiliaires observées, $\mathbf{x}_k^{\text{obs}}$, pour $k \in U$, tandis que μ_k et σ_k^2 sont des fonctions de $\mathbf{x}_k^{\text{obs}}$. Les estimateurs – asymptotiquement sans biais par rapport à m et convergents sous m – de μ_k et de σ_k^2 sont désignés par $\hat{\mu}_k$ et $\hat{\sigma}_k^2$, respectivement. Puisque nous conditionnerons systématiquement sur \mathbf{X}^{obs} , nous excluons cette notation pour simplifier. Par exemple, $E_m(y_k | \mathbf{X}^{\text{obs}})$ sera écrit sous la forme $E_m(y_k)$.

Dans le modèle (4.1), nous conditionnons sur les variables auxiliaires observées. Étant donné que le profil de non-réponse du vecteur \mathbf{x} n'est pas le même pour tous les non-répondants, nous devons valider et ajuster un modèle conditionnel distinct pour chaque profil de non-réponse. En principe, ces modèles conditionnels devraient être utilisés pour déterminer quelles méthodes d'imputation il convient de choisir. Notons que le modèle (4.1) se réduit au modèle conditionnel standard (par exemple, Särndal 1992) quand le vecteur \mathbf{x} de variables auxiliaires ne comprend pas de valeurs manquantes.

Remarque : Pour que la méthode d'estimation de la variance à la section 5 soit valide, il faut spécifier correctement μ_k et σ_k^2 . Bien qu'une forme paramétrique de μ_k soit souvent acceptable, il pourrait être plus difficile de déterminer une forme paramétrique appropriée de σ_k^2 . Pour éviter ce problème et pour obtenir une certaine robustesse aux erreurs de spécification de la variance du modèle, σ_k^2 peut être estimé de façon non paramétrique ; on trouvera dans l'étude empirique de Beaumont, Haziza et Bocci (2011) une illustration de cette propriété dans le contexte d'une imputation par valeur auxiliaire. Relativement à l'imputation par donneur, Beaumont et Bocci (2009) ont montré empiriquement que l'estimation non paramétrique de μ_k et σ_k^2 , par voie de lissage par splines pénalisées, réduit de façon significative la vulnérabilité de notre estimateur de la variance aux erreurs de spécification de la moyenne et de la variance du modèle.

En complément du modèle d'imputation (4.1), nous supposons que :

$$(4.2) \quad F(\mathbf{Y} | s, s_r, \mathbf{X}^{\text{obs}}, \mathbf{Z}, \mathbf{D}) = F(\mathbf{Y} | \mathbf{X}^{\text{obs}}),$$

et

$$\beta_1 = \sum_{k \in s_r^{(1)}} \omega_k^{(1)} y_k / \sum_{k \in s_r^{(1)}} \omega_k^{(1)} x_{1k}$$

$$\beta_2 = \sum_{k \in s_r^{(2)}} \omega_k^{(2)} y_k / \sum_{k \in s_r^{(2)}} \omega_k^{(2)}$$

respectivement. Dès lors, $\hat{\mu}_k = \beta_1 x_{1k}$, pour $k \in s_r^{(1)}$ ou $k \in s_m^{(1)}$, et $\hat{\mu}_k = \beta_2$, pour $k \in s_r^{(2)}$ ou $k \in s_m^{(2)}$. Tout comme à la section 2, on peut envisager l'utilisation de l'estimateur – potentiellement plus efficace – $\hat{\beta}_2^* = \sum_{k \in s_r} \omega_k^{(2)} y_k / \sum_{k \in s_r} \omega_k^{(2)}$ à la place de β_2 . Malheureusement, $\hat{\beta}_2^*$ est biaisé sous le modèle, puisque :

$$(4.3) \quad E_m(\hat{\beta}_2^* | s, s_r) = \beta_2 + \frac{\sum_{k \in s_r^{(1)}} \omega_k^{(1)} (x_{1k} \beta_1 - \beta_2)}{\sum_{k \in s_r^{(1)}} \omega_k^{(1)}}.$$

Tel que mentionné plus haut, si l'on suppose que les variables x_{1k} sont des variables aléatoires indépendamment distribuées de moyenne μ_x et de variance σ_x^2 , $\beta_2 = \beta_1 \mu_x$ et l'équation (4.3) peut être reformulée ainsi :

$$E_m(\hat{\beta}_2^* | s, s_r) = \beta_2$$

$$(4.4) \quad + \beta_1 \frac{\sum_{k \in s_r^{(1)}} \omega_k^{(1)} \sum_{k \in s_r^{(2)}} \omega_k^{(2)} (x_{1k} - \mu_x)}{\sum_{k \in s_r^{(1)}} \omega_k^{(1)} \sum_{k \in s_r^{(2)}} \omega_k^{(2)}}.$$

où $F(\cdot)$ dénote la fonction de répartition, \mathbf{Y} et \mathbf{D} sont des vecteurs à N éléments contenant respectivement y_k et d_k comme k^{e} élément, et \mathbf{Z} est une matrice de N lignes d'information sur le plan de sondage, qui contient implicitement ou explicitement l'information sur les probabilités de sélection π_k et les probabilités de sélection conjointe π_{kl} , pour $k, l \in U$. Cette hypothèse, souvent implicite dans d'autres articles, nous permet de traiter les indicateurs de réponse, les indicateurs de domaine et l'information sur le plan de sondage comme étant fixes lorsque nous considérons les espérances et les variances sous le modèle. Il faut choisir soigneusement les variables auxiliaires pour satisfaire à cette hypothèse. Par exemple, l'information sur le plan de sondage et les indicateurs de domaine devraient être envoyés à titre d'éventuelles variables auxiliaires.

La stratégie d'imputation exposée dans l'exemple amorcé à la section 2 pourrait être justifiée au moyen d'un modèle avec $\mu_k = \beta_1 x_{1k}$ et $\sigma_k^2 = \sigma_1^2 x_{1k}$, pour $k \in s_r^{(1)}$ ou $k \in s_m^{(1)}$, et $\mu_k = \beta_2$ et $\sigma_k^2 = \sigma_2^2$, pour $k \in s_r^{(2)}$ ou $k \in s_m^{(2)}$. Les paramètres β_1 , β_2 , σ_1^2 et σ_2^2 du modèle sont inconnus. Il est à noter que, si l'on suppose que les variables x_{1k} sont des variables aléatoires indépendamment distribuées de moyenne μ_x et de variance σ_x^2 , $\beta_2 = \beta_1 \mu_x$ et $\sigma_2^2 = \beta_1^2 \sigma_x^2 + \sigma_1^2 \mu_x$. On obtient les valeurs imputées $y_k^* = \hat{\mu}_k$, pour $k \in s_m$, en estimant les paramètres β_1 et β_2 du modèle à partir des données observées. Ainsi, les estimateurs sans biais par rapport à m de β_1 et β_2 pourraient être

par valeur auxiliaire. On trouvera un bon examen de ces méthodes dans Haziza (2009). Mentionnons que l'imputation par valeur auxiliaire ne fait pas appel aux valeurs de y des répondants, c'est-à-dire, $y_k^* = \phi_{(j)}^{0k}$ (voir Beaumont, Haziza et Bocci 2011). Dans le cas de l'imputation par donneur, la valeur imputée y_k^* est égale à la valeur de y d'un répondant choisi de façon appropriée (donneur), de sorte que $\phi_{(j)}^{0k} = 0$ et $\phi_{(j)}^{lk} = 0$ pour tous les répondants $l \in s_r$ sauf un. On trouvera des expressions détaillées de $\phi_{(j)}^{0k}$ and $\phi_{(j)}^{lk}$ dans le guide méthodologique du SEVANI (Beaumont, Bissounette et Bocci 2010), que l'on peut obtenir sur demande auprès des auteurs.

Supposons que $\Omega_{(j)}^l y_k^* = \sum_{k \in s_m^m} w_k d_k y_k^*$ est la contribution de la méthode d'imputation j à l'estimateur θ_l . La forme (3.1) permet de décomposer $\Omega_{(j)}^l$ de la façon suivante :

$$\Omega_{(j)}^l y_k^* = \sum_{k \in s_m^m} w_k d_k y_k^* = \sum_{k \in s_m^m} w_k d_k \phi_{(j)}^{0k} + \sum_{l \in s_r} w_l \sum_{k \in s_m^m} w_k d_k \phi_{(j)}^{lk} = W_{(j)}^{0d} + \sum_{l \in s_r} W_{(j)}^{dl} y_l, \quad (3.2)$$

où $W_{(j)}^{0d} = \sum_{k \in s_m^m} w_k d_k \phi_{(j)}^{0k}$ et $W_{(j)}^{dl} = \sum_{k \in s_m^m} w_k d_k \phi_{(j)}^{lk}$. À partir de (3.2), l'estimateur imputé (2.1) peut être exprimé sous la forme linéaire suivante :

$$\theta_l = \sum_{j=1}^J \Omega_{(j)}^l y_k^* + \sum_{k \in s_r} W_{(j)}^{0d} + \sum_{k \in s_r} \left(W_{(j)}^{dk} + W_{(j)}^{dk} \right) y_k, \quad (3.3)$$

où $W_{(j)}^{0d} = \sum_{j=1}^J W_{(j)}^{0d}$ et $W_{(j)}^{dk} = \sum_{j=1}^J W_{(j)}^{dk}$.

Si l'on reprend l'exemple présenté à la fin de la section 2, on constate que, aux fins d'imputation par le ratio, $\phi_{(1)}^{0k} = 0$ et $\phi_{(1)}^{lk} = \omega_{(1)}^{lk} / \sum_{l \in s_r} \omega_{(1)}^{lk} x_{1l}$, pour $l \in s_r$, où $\omega_{(1)}^{lk} = 0$ pour $l \in s_r^r$. Dans le cas de l'imputation par la moyenne, $\phi_{(2)}^{0k} = 0$ et $\phi_{(2)}^{lk} = \omega_{(2)}^{lk} / \sum_{l \in s_r} \omega_{(2)}^{lk}$, pour $l \in s_r$. Par conséquent, $W_{(1)}^{0d} = 0$, $W_{(2)}^{0d} = 0$, $W_{(1)}^{dk} = \omega_{(1)}^{dk} / \sum_{k \in s_m^m} w_k d_k x_{1k}$ et $W_{(2)}^{dk} = \omega_{(2)}^{dk} / \sum_{k \in s_m^m} w_k d_k$.

et $W_{(2)}^{dk} = \omega_{(2)}^{dk} / \sum_{k \in s_m^m} w_k d_k$. Cela signifie que $W_{(+) }^{0d} = 0$ et que $W_{(+) }^{dk} = W_{(1)}^{dk} + W_{(2)}^{dk}$.

4. Approche d'inférence et principales hypothèses

Nous considérons trois sources de variabilité quand nous évaluons les espérances et les variances de l'estimateur imputé : la variabilité due au modèle d'imputation, au plan d'échantillonnage et au mécanisme de non-réponse. Il est à noter que l'utilisation d'un modèle d'imputation pour faire une inférence lorsqu'il y a imputation est mentionnée par Rubin (1987), Hidiroglou (1989) et Särndal (1992). Dans la

L'ensemble de non-répondants s_m est subdivisé en deux sous-ensembles, $s_{(1)}^m$ et $s_{(2)}^m$, d'après la disponibilité de x_{1l} .

De même, l'ensemble de répondants est subdivisé entre $s_{(1)}^r$ et $s_{(2)}^r$. Dans cet exemple, nous pourrions utiliser l'imputation par le ratio pour imputer les valeurs manquantes de y dans $s_{(1)}^m$ et l'imputation par la moyenne pour les imputer dans $s_{(2)}^m$. Il faut mentionner que l'on pourrait opter pour l'imputation par régression linéaire simple plutôt que pour l'imputation par le ratio (si elle est mieux ajustée aux données). Nous avons opté ici pour l'imputation par le ratio parce qu'il s'agit d'une méthode simple et qui est souvent utilisée dans les enquêtes auprès des entreprises.

Seuls les répondants du sous-ensemble $s_{(1)}^r$ peuvent être utilisés pour imputer les valeurs manquantes de y dans $s_{(1)}^m$ avec l'imputation par le ratio. La valeur imputée pour une unité k dans $s_{(1)}^m$ est $y_k^* = x_{1k} \sum_{l \in s_{(1)}^r} \omega_{(1)}^{lk} / \sum_{l \in s_{(1)}^r} \omega_{(1)}^{lk} x_{1l}$, où $\omega_{(1)}^{lk}$ est un poids utilisé aux fins de l'imputation par le ratio (méthode 1). Les choix typiques seront $\omega_{(1)}^{lk} = w_l$ (imputation pondérée par les poids de sondage) ou $\omega_{(1)}^{lk} = 1$ (imputation non pondérée). Dans le cas de l'imputation par la moyenne, on peut utiliser les répondants du sous-ensemble $s_{(2)}^r$ ainsi que ceux du sous-ensemble $s_{(1)}^r$ pour imputer les valeurs manquantes de y dans $s_{(2)}^m$. Dans la pratique, il est courant d'utiliser les deux ensembles de répondants afin d'accroître la stabilité de la moyenne imputée. La valeur imputée pour une unité k dans $s_{(2)}^m$ est

$$y_k^* = \sum_{l \in s_r} \omega_{(2)}^{lk} y_{1l} / \sum_{l \in s_r} \omega_{(2)}^{lk},$$

où $\omega_{(2)}^{lk}$ est un poids utilisé aux fins de l'imputation par la moyenne (méthode 2) (les choix typiques pour $\omega_{(2)}^{lk}$ seront les mêmes que pour $\omega_{(1)}^{lk}$, soit $\omega_{(2)}^{lk} = w_l$ ou $\omega_{(2)}^{lk} = 1$). Cela signifie que les unités faisant partie de $s_{(1)}^r$ peuvent être utilisées dans les deux méthodes. Cette situation soulève des problèmes lorsque l'on veut estimer la variance associée à l'estimateur par imputation composite ainsi obtenu. Ces problèmes sont commentés à la section 5.

3. Qu'est-ce que l'imputation linéaire ?

La méthode d'imputation j est dite linéaire si la valeur imputée y_k^* pour une unité échantillonnée $k \in s_{(j)}^m$ peut s'écrire sous la forme linéaire :

$$y_k^* = \phi_{(j)}^{0k} + \sum_{l \in s_r} \phi_{(j)}^{lk} y_{1l}. \quad (3.1)$$

Les quantités $\phi_{(j)}^{0k}$ et $\phi_{(j)}^{lk}$, pour $l \in s_r$, sont obtenues sans que l'on utilise les valeurs de y , mais elles peuvent dépendre de s et s_r . Dans la pratique, plusieurs des méthodes d'imputation les plus courantes satisfont à la forme linéaire (3.1), par exemple l'imputation par régression linéaire (pondérée ou non), l'imputation par donneur et l'imputation

valeurs manquantes dans différents sous-ensembles. Nous pouvons exprimer de la façon suivante l'estimateur imputé ainsi obtenu :

$$\hat{\theta}_I = \sum_{k \in s_m} w_k d_k y_k + \sum_{k \in s_r} w_k d_k y_k^* \quad (2.1)$$

$$= \sum_{k \in s_r} w_k d_k y_k + \sum_{j=1}^J \sum_{k \in s_m^{(j)}} w_k d_k y_k^* \quad (2.1)$$

où y_k^* est la valeur imputée de y pour l'unité k .

L'imputation composite est couramment utilisée dans les enquêtes auprès des entreprises. La raison en est qu'il y a des valeurs manquantes dans les variables auxiliaires dont on se sert pour l'imputation. De manière à préciser les idées, représentons au moyen de x_k le vecteur complet de variables auxiliaires pour l'unité k . Idéalement, toutes les valeurs de y qui sont manquantes seraient imputées au moyen d'une seule méthode d'imputation à partir du vecteur complet x_k . Malheureusement, certaines valeurs peuvent manquer dans les variables auxiliaires, de sorte que nous ne pouvons nous servir de x_k pour imputer les valeurs de y manquantes pour certains non-répondants ; nous pourrions uniquement utiliser un sous-ensemble de x_k . Le vecteur des variables auxiliaires observées pour l'unité k est désigné par $x_{k,obs}$. Ce vecteur ne contient pas nécessairement les mêmes variables observées d'une unité à l'autre. Aux fins d'imputer les valeurs de y manquantes pour une unité k donnée, une méthode d'imputation est choisie en fonction des variables auxiliaires disponibles. Puisqu'il peut exister un certain nombre de profils de non-réponse à l'intérieur du vecteur complet de variables auxiliaires, la stratégie d'imputation peut comporter un certain nombre de méthodes d'imputation.

Exemple

L'exemple qui suit pourra aider à mieux saisir les questions soulevées par l'estimation de la variance en cas d'imputation composite. Supposons que le vecteur complet de variables auxiliaires pour l'unité k est $x_k = (x_{1k}, x_{2k})$, où x_{1k} est fortement corrélé à y_k mais peut présenter des valeurs manquantes, et x_{2k} est une constante pour toutes les unités échantillonnées ($x_{2k} = 1, k \in s$). Idéalement, on utilisera x_{1k} pour imputer y_k si cette valeur est manquante. Si l'on ne dispose pas de x_{1k} , on peut uniquement utiliser x_{2k} . Le tableau 1 résume l'information disponible pour les divers sous-ensembles de l'échantillon s .

Tableau 1
Information disponible quand il existe une variable auxiliaire x_1 et une constante x_2

Sous-ensembles				
y	x_1	x_2	x_{obs}	
$s_r^{(1)}$	O	O	O	(x_1, x_2)
$s_r^{(2)}$	O	O	O	(M, x_2)
$s_m^{(1)}$	M	O	O	(x_1, x_2)
$s_m^{(2)}$	M	M	O	(M, x_2)

O : valeur observée ; M : valeur manquante.

2. Qu'est-ce que l'imputation composite ?

Supposons que nous voulons estimer le total d'un domaine de population $\theta = \sum_{k \in U} d_k y_k$, où U est la population finie de taille N , y est la variable d'intérêt et d est une variable indicatrice de domaine précisant si l'unité de population k fait partie du domaine d'intérêt ($d_k = 1$ ou non ($d_k = 0$). Un échantillon s de taille n est choisi à partir de la population finie U selon un plan d'échantillonnage probabiliste $p(s)$. S'il n'y a pas de valeurs manquantes, θ peut être estimé au moyen de l'estimateur de Horvitz-Thompson $\hat{\theta} = \sum_{k \in s} w_k d_k y_k$, où $w_k = 1/\pi_k$ est le poids de sondage et π_k est la probabilité de sélection de l'unité k . Il serait possible d'étendre nos résultats aux estimateurs de calage, mais, par souci de simplicité, nous ne le faisons pas ici.

La variable y peut être manquante pour certaines des unités échantillonnées, mais nous faisons l'hypothèse que la variable indicatrice de domaine d est toujours observée pour ces unités. L'ensemble des unités échantillonnées qui sont assorties d'une valeur de y observée – les répondants – est désigné au moyen de s_r . On suppose que cet ensemble a été généré conformément à un mécanisme de non-réponse $q(s_r | s)$. L'ensemble des non-répondants est désigné par $s_m = s - s_r$. Il est subdivisé en J sous-ensembles mutuellement exclusifs, $s_m^{(j)}, j = 1, \dots, J$, de sorte que $s_m = \bigcup_{j=1}^J s_m^{(j)}$, si l'on a recours à l'imputation composite au moyen de $J > 1$ méthodes d'imputation. Toutes les valeurs de y manquantes dans un sous-ensemble donné $s_m^{(j)}$ sont imputées au moyen de la même méthode, j . Toutefois, différentes méthodes d'imputation sont utilisées pour imputer les

Estimation de la variance sous imputation composite : méthodologie programmée dans le SEVANI

Jean-François Beaumont et Joël Bissonnette¹

Résumé

L'imputation composite est fréquemment employée dans les enquêtes auprès des entreprises. Le terme « composite » signifie que l'on utilise plus d'une méthode d'imputation pour remplacer les valeurs manquantes d'une variable d'intérêt. Le choix de la méthode dépendra de la disponibilité de variables auxiliaires. Par exemple, on pourra utiliser l'imputation par le ratio pour imputer une valeur manquante si l'on dispose d'une variable auxiliaire ; sinon, on pourra opter pour l'imputation par la moyenne.

Il y a une abondante littérature consacrée au problème que pose l'estimation de la variance lorsque l'on utilise une méthode d'imputation unique, dont les excellents comptes rendus de Lee, Rancourt et Särndal (2001) ainsi que de Haziza (2009). Par contre, on recense peu de travaux sur l'estimation de la variance lorsque l'on a recours à l'imputation composite, même si ce type d'imputation est souvent utilisé dans la pratique. Mentionnons notamment Rancourt, Lee et Särndal (1993), qui ont proposé et évalué empiriquement un estimateur de variance jackknife et par linéarisation (1997) ont poussé plus loin l'étude théorique et ont obtenu des estimateurs de variance jackknife et par linéarisation convergents par rapport au plan de sondage. Dans les deux articles, les auteurs ont considéré deux méthodes d'imputation, dont l'imputation par le ratio, dans des conditions d'échantillonnage aléatoire simple où l'on supposait que la non-réponse était uniforme. Puis, Felix et Rancourt (2001) ont étendu la méthode générale proposée par Särndal (1992)

Mots clés : Imputation par valeur auxiliaire ; imputation composite ; imputation par donneur ; modèle d'imputation ; imputation linéaire ; imputation par régression ; SEVANI.

1. Introduction

et par Deville et Särndal (1994) à l'imputation composite au moyen d'hypothèses simplifiées. Enfin, pour tenir compte de l'imputation composite, Shao et Steel (1999) ont élaboré une approche inverse générale d'estimation de la variance qui est fort intéressante (voir également Kim et Rao 2009). Shao et Steel (1999) ont fait valoir que leur approche inverse donnait lieu à des calculs moins compliqués que celle de Deville et Särndal (1994). Nous ne partageons pas entièrement ce point de vue. Nos résultats montrent que, de façon générale, l'application que nous faisons de l'approche de Särndal donne en fait lieu à des calculs plus simples que ce n'est le cas avec celle de Shao et Steel. Cela dit, l'approche inverse pourrait devenir beaucoup plus attrayante lorsque la fraction de sondage est négligeable et que l'on opte pour une technique d'estimation de la variance par réplication (on trouvera à la section 7 des commentaires plus détaillés à ce sujet).

Nous utilisons comme point de départ la méthode proposée par Särndal (1992). Elle requiert un modèle d'imputation valide, c'est-à-dire un modèle pour la variable qui est imputée. À première vue, l'extension de cette méthode à l'imputation composite semble assez fastidieuse, comme l'ont souligné Shao et Steel (1999), jusqu'à ce que l'on remarque que la plupart des méthodes d'imputation utilisées dans la pratique donnent des valeurs observées de la variable d'intérêt. Cela simplifie considérablement le calcul d'un estimateur de variance, même si l'on n'utilise qu'une seule méthode d'imputation. Pour estimer la part de la variance

1. Jean-François Beaumont, Statistique Canada, Division de la recherche et de l'innovation en statistique, pré Tunney, Ottawa (Ontario), Canada, K1A 0T6. Courriel : jean-francois.beaumont@statcan.gc.ca ; Joël Bissonnette, Statistique Canada, Division des méthodes d'enquêtes auprès des entreprises, pré Tunney, Ottawa (Ontario), Canada, K1A 0T6. Courriel : joel.bissonnette@statcan.gc.ca.

(2007), le but ultime de l'estimation sur petits domaines est de faire des inférences au sujet des caractéristiques des petits domaines conditionnellement aux valeurs réalisées (mais inconnues) des effets de petit domaine, c'est-à-dire par rapport à (1). On peut considérer cela comme étant un objectif fondé sur le plan (comme dans Longford 2007) ou, comme nous le préférons, un objectif fondé sur un modèle qui ne rentre pas vraiment dans le cadre habituel des effets aléatoires pour l'estimation sur petits domaines. Dans l'un et l'autre cas, nous nous intéressons à la variabilité par rapport à des valeurs prévues fixes propres au domaine, ce qui est conforme au concept de variabilité qui est habituellement appliqué en inférence au niveau de la population.

Remerciements

Les auteurs remercient le rédacteur en chef, le rédacteur associé et deux examinateurs de leurs commentaires et suggestions précieux qui leur ont permis d'améliorer considérablement l'article.

Bibliographie

Chambers, R.L., et Dunstan, R. (1986). Estimating distribution functions from survey data. *Biometrika*, 73, 597-604.

Chambers, R., et Tzavidis, N. (2006). M-quantile models for small area estimation. *Biometrika*, 93, 255-268.

Chandra, H., et Chambers, R. (2009). Multipurpose weighing for small area estimation. *Journal of Official Statistics*, 25(3), 379-395.

Chandra, H., Chambers, R. et Salvati, N. (2011). Small area estimation of proportions in business surveys. Paraîtra dans le *Journal of Statistical Computation and Simulation*.

Henderson, C.R. (1953). Estimation of variance and covariance components. *Biometrics*, 9, 226-252.

Longford, N.T. (2007). De l'erreur-type des estimateurs pour petits domaines fondés sur un modèle. *Techniques d'enquête*, 33, 81-92.

Opsomer, J.D., Claeskens, G., Ranalli, M.G., Kaetmann, G. et Breidt, F.J. (2008). Nonparametric small area estimation using penalized spline regression. *Journal of the Royal Statistical Society, Séries B*, 70, 265-286.

Prasad, N.G.N., et Rao, J.N.K. (1990). The estimation of the mean squared error of small area estimators. *Journal of the American Statistical Association*, 85, 163-171.

Rao, J.N.K. (2003). *Small Area Estimation*. New York : John Wiley & Sons, Inc.

Royal, R.M. (1976). The linear least squares prediction approach to two-stage sampling. *Journal of the American Statistical Association*, 71, 657-664.

Royal, R.M., et Cumberland, W.G. (1978). Variance estimation in finite population sampling. *Journal of the American Statistical Association*, 73, 351-358.

Salvati, N., Chandra, H., Ranalli, M.G. et Chambers, R. (2010). Small area estimation using a nonparametric model based direct estimator. *Computational Statistics and Data Analysis*, 54, 2159-2171.

Salvati, N., Tzavidis, N., Pratesi, M. et Chambers, R. (2011). Small area estimation via M-quantile geographically weighted regression. À venir dans *TEST*, DOI 10.1007/s11749-010-0231-1.

Sinha, S.K., et Rao, J.N.K. (2009). Robust small area estimation. *The Canadian Journal of Statistics*, 37(3), 381-399.

Street, J.O., Carroll, R.J. et Ruppert, D. (1988). A note on computing robust regression estimates via iteratively reweighted least squares. *The American Statistician*, 42, 152-154.

Tzavidis, N., Marchetti, S. et Chambers, R. (2010). Robust prediction of small area means and distributions. *Australian and New Zealand Journal of Statistics*, 52, 167-186.

Valliant, R., Dorfman, A.H. et Royall, R.M. (2000). *Finite Population Sampling and Inference*. New York : John Wiley & Sons, Inc.

4. Conclusion et discussion

demandant beaucoup de calculs. Notons toutefois que, pour des tailles d'échantillon de domaine très petites, l'estimateur de l'EQM robuste au biais proposé dans le présent article reste instable.

Une future piste de recherche pourrait consister à comparer la méthode analytique d'estimation de l'EQM proposée ici aux estimateurs de l'EQM fondés sur le bootstrap, par exemple l'estimateur bootstrap non paramétrique de l'EQM de l'estimateur du M-quantile proposé par Tzavidis, Marchetti et Chambers (2010), et l'estimateur bootstrap de l'EQM pour l'estimateur EBLUP robuste proposé par Sinha et Rao (2009). Une question clé dans cette étude sera de déterminer si d'autres options d'estimateur bootstrap de l'EQM sont plus stables, surtout pour les petites tailles d'échantillon de domaine.

L'extension de l'approche d'estimation conditionnelle de l'EQM aux situations non linéaires d'estimation sur petits domaines reste à faire. Cependant, puisque cette approche est étroitement reliée à l'estimation robuste de l'EQM en population fondée sur la linéarisation par développement en série de Taylor (ainsi que l'estimation jackknife de l'EQM, voir Valliant, Dorfman et Royall 2000, section 5.4.2), il devrait être possible de développer des extensions appropriées pour les méthodes d'estimation non linéaires sur petits domaines correspondantes. Bien que nous ne présentions pas les résultats pertinents ici, nous voyons pour preuve le fait que la méthode d'estimation conditionnelle de l'EQM décrite dans le présent article a déjà été utilisée pour estimer l'EQM de l'estimateur EDFM quand il est appliqué à des variables qui ne se prêtent pas à la modélisation linéaire mixte, c'est-à-dire celles pour lesquelles la proportion de valeurs nulles est élevée (Chandra et Chambers 2009), et à des variables catégoriques (Chandra, Chambers et Salva 2011). Plus récemment, l'approche a également été utilisée pour estimer l'EQM des estimateurs sur petits domaines des M-quantile géographiquement pondérés dans des situations où les valeurs de petit domaine sont spatialement corrélées (Salva, Tzavidis, Pratesi et Chambers 2011). Il a également été employé par Salva, Chandra, Ranalli et Chambers (2010) pour estimer l'EQM d'estimateurs sur petits domaines fondés sur un modèle de petit domaine non paramétrique (Opsomer, Claeskens, Ranalli, Kauermann et Breidt 2008).

Comme le montre clairement le présent exposé, notre approche prêterait d'estimation de l'EQM repose sur l'hypothèse que l'EQM qui nous intéresse vraiment est celle définie par le modèle propre au domaine (1). Cette approche diffère de celle habituellement suivie pour définir l'EQM dans l'estimation sur petits domaines, c'est-à-dire en adoptant un concept de moyenne de l'EQM sur les domaines comme mesure appropriée de l'exactitude d'un estimateur sur petits domaines. Comme l'a fait remarquer Longford

Dans le présent article, nous proposons une méthode robuste au biais et facile à appliquer d'estimation de l'EQM conditionnelle des estimateurs pseudo-linéaires de moyennes (et de totaux) de petits domaines. Nos résultats empiriques montrent que cette méthode d'estimation de l'EQM donne des résultats raisonnablement bons en ce qui concerne le biais quand elle est utilisée pour estimer l'EQM sous un modèle ainsi que sous un plan de sondage des trois estimateurs pseudo-linéaires assez différents pris en considération dans la présente étude. Cependant, ces résultats améliorés pour ce qui est du biais ont pour prix une augmentation de la variabilité. En particulier, quand les tailles d'échantillon de domaine sont très petites, nous ne recommandons pas d'appliquer la méthode d'estimation de l'EQM que nous proposons à un estimateur conditionnellement biaisé tel que l'EBLUP.

L'usage de l'EBLUP est très répandu pour l'estimation sur petits domaines et, dans ce contexte, l'estimateur de l'EQM dépendant d'un modèle PR0 proposé pour l'EBLUP par Prasad et Rao (1990) est sans biais quand les hypothèses qui sous-tendent ce modèle sont valides (SIM1-A/B et SIM2-A/B dans nos simulations fondées sur un modèle), mais il contient un biais en présence d'effets de domaine aberrants (SIM3-A/A* et SIM3-B/B*). Il s'agit aussi de l'estimateur de l'EQM le plus stable dans les simulations fondées sur un modèle. Toutefois, étant donné sa construction fondée sur la moyenne de domaines, il ne détecte pas l'EQM propre au domaine de l'EBLUP dans nos deux simulations fondées sur un plan de sondage, où nous pouvons seulement considérer que le modèle linéaire mixte suppose n'est qu'une approximation correcte. Cela donne à penser que la méthode d'estimation conditionnelle de l'EQM que nous proposons devrait être envisagée comme une alternative au modèle PR0 dans les situations où il existe un doute quant à l'exactitude de la spécification du modèle mixte linéaire au niveau du petit domaine ou dans celles où les tailles d'échantillon de domaine ne sont pas faibles. Les résultats empiriques présentés ici permettent de se faire une idée de ce qui constitue une petite taille d'échantillon.

Si l'utilisateur doute de la validité du modèle mixte linéaire supposé, il pourrait envisager l'estimation fondée sur un modèle de rechange d'application plus générale, par exemple le modèle du M-quantile, ou remplacer l'EBLUP par une option plus robuste aux valeurs aberrantes (Sinha et Rao 2009). Dans le premier cas, l'approche que nous proposons ici est, à l'heure actuelle, la seule approche analytique de l'estimation de l'EQM, tandis que dans le second cas, l'approche proposée fournit une option analytique pour remplacer les méthodes bootstrap d'estimation de l'EQM

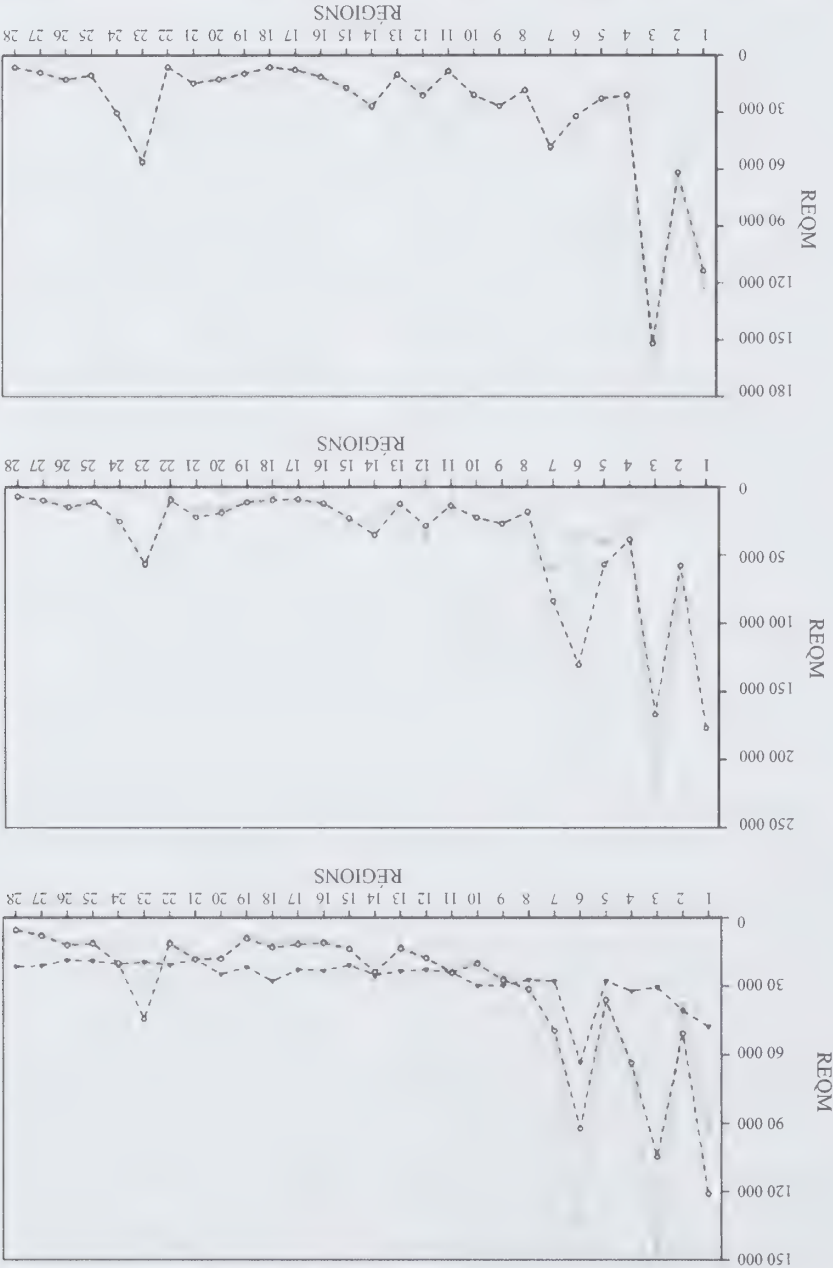


Figure 3 Valeurs au niveau de la r  ion de la REQM sous le plan r  el (trait plein) et de la REQM estim  e moyenne (trait interrompu) obtenues dans les simulations fond  es sur le plan en utilisant la population de fermes de l'AAGIS. Les r  ions sont class  es par ordre de taille de population croissante. Les valeurs pour l'estimateur PR0 sont indiqu  es par Δ, tandis que celles pour l'estimateur conditionnel sont indiqu  es par o. Les graphiques montrent les r  sultats pour les estimateurs EBLUP (en haut), EDFM (au centre) et M-quantile (en bas)

Tableau 7 Propri  t  s des estimateurs EBLUP et de l'EQM en pr  sence de domaines dont l'  chantillon est nul

M��thode de pond��ration/Estimateur					AAGIS, n_i m��dian = 9				
SIM1-A, n_i m��dian = 10					BR(m)				
EQM Estimateur					REQMR(M)				
BR(M)					BR(M)				
REQMR(M)					REQMR(M)				
Domaines avec $n_i > 0$					2,29				
(13)/EBLUP					24,94				
(23)/EBLUP synth��tique					87,45				
Domaines avec $n_i = 0$					0,00				
BR(M)					0,52				
REQMR(M)					1,25				
Domaines avec $n_i > 0$					29,91				
(13)/PR0					23,87				
(23)/PR0					29,07				
Domaines avec $n_i = 0$					0,5				
REQMR(M)					11				
(13)/PR0					50				
(23)/PR0					35				
Domaines avec $n_i = 0$					-3,6				
(23)/Conditionnel					-31,45				
(23)/Conditionnel					101				

L'importance de cette hypothèse en comparant les propriétés de l'estimation de l'EBLUP et de l'estimation de son EQM pour les domaines échantillonnés à celles observées pour l'EBLUB synthétique pour les domaines pour lesquels aucune donnée d'échantillon n'est disponible. Nous présentons deux situations. La première est une modification de la simulation SIM1-A fondée sur un modèle avec une petite taille moyenne d'échantillon et avec cinq domaines dont l'échantillon est nul. La deuxième est une modification similaire, en prenant une petite taille d'échantillon, de la simulation fondée sur un plan de sondage appliquée à la population de l'AAGIS, avec quatre domaines dont l'échantillon est nul. Il est évident que, quand le modèle qui

sous-tend l'EBLUP est effectivement vérifié (c'est-à-dire SIM1-A), l'estimation ainsi que l'estimation de son EQM (fondée sur le modèle PR0 ou sur l'option conditionnelle) donne de bons résultats. Le problème tient au fait que, s'il existe un doute quant à la façon dont ce modèle tient (comme dans la population de l'AAGIS), l'EBLUP peut échouer et notre estimateur de son EQM peut aussi ne pas arriver à déceler ce problème. Cela est bien illustré par les résultats pour la population de l'AAGIS présentés au tableau 7, où nous voyons que l'estimateur PR0 ainsi que l'estimateur conditionnel de l'EQM pour l'EBLUP synthétique n'arrivent pas du tout à déceler l'important biais positif de ce dernier et présentent donc un grand biais vers le bas.

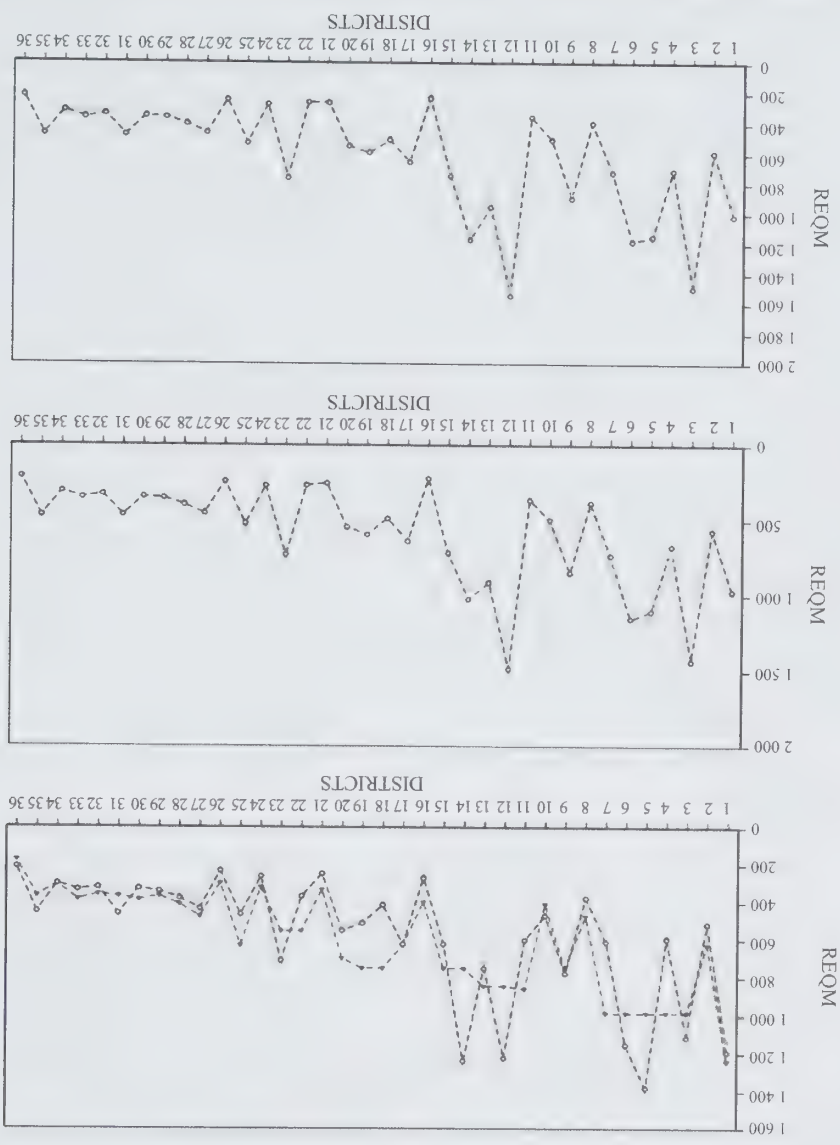


Figure 2 Valeurs au niveau du district de la REQM sous le plan réel (trait plein) et de la REQM estimée moyenne (trait interrompu) obtenues dans les simulations fondées sur le plan en utilisant la population de ménages de l'Albanie. Les districts sont classés par ordre de taille de population croissante. Les valeurs pour l'estimateur PR0 sont indiquées par Δ, tandis que celles pour l'estimateur conditionnel sont indiquées par o. Les graphiques montrent les résultats pour les estimateurs EBLUP (en haut), EDFM (au centre) et M-quantile (en bas)

Tableau 5

Propriétés des estimateurs des moyennes régionales et des estimateurs de leur EQM – Population de ménages de l'Albanie

Méthode de pondération					
n_i médian = 56					
n_i médian = 9					
Estimateur	BR(m)	REQMR(m)	BR(m)	REQMR(m)	REQMR(m)
Régession	0,04	6,25	-0,13	16,56	16,56
EBLUP, (13)	0,42	5,90	1,62	12,42	12,42
EDFM, (18)	0,03	6,14	0,04	16,92	16,92
M-quantile, (22)	0,04	6,07	-0,05	16,60	16,60
Méthode/EQM	BR(M)	REQMR(M)	BR(M)	REQMR(M)	REQMR(M)
Régession /VReg	17,6	42	11,2	42	42
EBLUP/PR0	14,6	44	10,5	50	50
EBLUP/PR1	14,4	43	8,8	48	48
EBLUP/PR2	14,5	43	9,7	48	48
EBLUP/Cconditionnel	0,1	24	7,7	99	99
EDFM/Cconditionnel	-0,8	25	-5,5	64	64
M-quantile/Cconditionnel	2,9	27	-2,0	75	75

Tableau 6

Propriétés des estimateurs des moyennes régionales et des estimateurs de leur EQM – Population de fermes de l'AAGIS

Méthode de pondération					
n_i médian = 53					
n_i médian = 8					
Estimateur	BR(m)	REQMR(m)	BR(m)	REQMR(m)	REQMR(m)
Régession	0,03	13,36	0,08	29,83	29,83
EBLUP, (13)	1,64	13,53	0,92	25,82	25,82
EDFM, (18)	-0,73	14,26	-1,02	37,77	37,77
M-quantile, (22)	-0,04	11,68	-0,15	32,22	32,22
Méthode/EQM	BR(M)	REQMR(M)	BR(M)	REQMR(M)	REQMR(M)
Régession /VReg	74,1	406	54,7	867	867
EBLUP/PR0	22,4	131	17,7	374	374
EBLUP/PR1	19,5	137	19,0	367	367
EBLUP/PR2	21,0	123	31,1	444	444
EBLUP/Cconditionnel	5,5	132	17,8	255	255
EDFM/Cconditionnel	-0,5	181	0,9	318	318
M-quantile/Cconditionnel	-0,7	69	-1,9	212	212

L'estimateur de l'EQM de l'estimateur par la régression présente un biais modéré ou élevé pour les deux populations et tous les scénarios de simulation. Dans le cas de la population albanaïenne, il semble soutenir la concurrence des autres estimateurs de l'EQM en ce qui concerne la REQMR, mais dans le cas de la population de l'AAGIS, il est clairement moins stable que les autres estimateurs de l'EQM. Enfin, l'estimateur conditionnel de l'EQM de l'estimateur du M-quantile donne de bons résultats, son biais relatif étant faible et sa stabilité, bonne, pour tous les scénarios de simulation et les deux populations, sauf dans le cas de la population albanaïenne avec une taille médiane d'échantillon $n_i = 9$ auquel cas la REQMR est de 75 %.

L'examen des valeurs de la REQMR propre au domaine présentées à la figure 2 pour la population albanaïenne et à la figure 3 pour la population de l'AAGIS donne une idée des raisons de ces différences de comportement. Notons que, dans les deux cas, les tailles d'échantillon sont celles tirées des enquêtes originales. Donc, à la figure 2, nous constatons que les estimateurs conditionnels de l'EQM suivent tous trois exceptionnellement bien les REQMR sous le plan propre aux districts de leurs estimateurs respectifs. En revanche, l'estimateur PR0 ne semble pas capable de saisir les différences entre districts de la REQMR sous le plan de l'estimateur EBLUP. À la figure 3, nous voyons que

L'estimateur conditionnel de l'EQM de l'estimateur du M-quantile donne de très bons résultats dans toutes les régions, tandis que l'estimateur correspondant de l'EQM de l'estimateur EDFM donne aussi de bons résultats dans toutes les régions sauf une (région 6), dans laquelle il surestime considérablement la REQMR sous le plan de ce prédicteur. Cette région mérite d'être mentionnée, parce que les échantillons qui sont déséquilibrés en ce qui a trait à la superficie dans cette région produisent des poids négatifs sous le modèle mixte linéaire supposé. Le tableau se complique lorsque l'on considère les propriétés de l'estimation de la REQMR propre à la région pour l'EBLUP à la figure 3. Ici, nous voyons clairement que l'estimateur conditionnel de l'EQM de l'EBLUP suit mieux la REQMR sous le plan propre à la région de ce prédicteur que l'estimateur PR0 de l'EQM. Sauf dans le cas de la région 6 (pour laquelle l'équilibre de l'échantillon pose problème), la variation régionale de la valeur de l'estimateur PR0 de la REQMR de l'EBLUP semble faible, ce qui indique un problème de biais sérieux.

Comme nous l'avons mentionné plus haut, il n'est pas rare de vouloir produire une estimation pour un petit domaine pour lequel il n'existe pas d'échantillon. Le cas échéant, on doit s'appuyer entièrement sur l'exactitude de la spécification du modèle. Au tableau 7, nous illustrons

populations, il existe des estimateurs indirects dont les propriétés sont un peu meilleures. Les simulations fondées sur un plan de sondage s'appuyant sur les populations albanienne et de l'AAGIS ont également été exécutées en prenant des tailles d'échantillon de domaine plus petites que celles utilisées dans les enquêtes originales. En particulier, pour la population albanaïenne, la taille globale d'échantillon a été réduite à $n = 291$ (avec une taille d'échantillon de district médiane de 9). De même, pour la population de l'AAGIS, la taille globale d'échantillon a été réduite à $n = 233$ (avec une taille médiane d'échantillon régional de 8). Comme prévu, la REQM des estimateurs ponctuels augmente à mesure que les tailles d'échantillon de domaine diminuent. Dans l'ensemble, sous ces tailles d'échantillon plus petites, la REQM de l'EBLUP s'améliore comparativement à celle des autres estimateurs. Cependant, puisqu'il est douteux que ces plans à taille d'échantillon réduite soient raisonnables, nous n'accordons pas trop d'importance aux résultats qui en découlent, nous contenant seulement de mentionner qu'ils sont utiles pour évaluer la performance des estimateurs de l'EQM avec des données réelles et de très petites tailles d'échantillon.

Si nous nous penchons sur les résultats de simulation obtenus en utilisant les tailles originales d'échantillons régionaux, nous voyons que, pour l'EBLUP, les trois estimateurs PR de l'EQM présentent un biais vers le haut important dans les deux ensembles de simulations fondées sur un plan de sondage ainsi qu'une instabilité plus grande (population albanaïenne, tableau 5) que celle des estimateurs conditionnels de l'EQM, ou comparable à celle-ci (population de l'AAGIS, tableau 6). Pour la population albanaïenne, les trois versions de l'estimateur conditionnel de l'EQM sont essentiellement sans biais, tandis que pour la population de l'AAGIS, elles présentent un biais faible ou modéré. Il convient de souligner que, pour la population albanaïenne (tableau 5), les propriétés relatives des estimateurs PR de l'EQM s'améliorent quand les échantillons deviennent plus petits. Cependant, cela tient au fait que les estimateurs conditionnels de l'EQM deviennent alors plus instables. Pour ces très petits échantillons de domaine, l'estimateur conditionnel de l'EQM comporte un biais moins important que l'estimateur PR de l'EQM (7,7 % vs 10,5 %), mais il est également plus instable (REQMR de l'estimateur conditionnel de l'EQM égal à 99 % vs 50 % pour l'estimateur PR de l'EQM). Néanmoins, ce n'est pas le cas pour la population de l'AAGIS, avec une taille médiane $n_i = 8$. Dans ce cas, les estimateurs PR de l'EQM donnent de mauvais résultats, les estimateurs conditionnels de l'EQM étant à la fois moins biaisés et plus stables.

étant définies par les régions et les tailles d'échantillon de strate étant définies par celles de l'échantillon original de l'AAGIS. Ces tailles d'échantillon varient de 6 à 117, avec une taille d'échantillon de région médiane de 53. Ici, Y représente le coût déboursé total (CDT) associé à l'exploitation de la ferme, et X est un vecteur qui comprend la superficie de la ferme (superficie), les effets pour six poststrates définies par trois zones climatiques et deux tranches de taille de ferme, ainsi que l'interaction de ces variables. Dans l'échantillon original de l'AAGIS, la relation entre le CDT et la superficie varie de manière significative entre les six poststrates, la valeur globale du R-carré étant d'environ 0,46 après la suppression de deux valeurs aberrantes. Par conséquent, dans le modèle de prédiction, les effets fixes ont été spécifiés comme correspondant à un ajustement linéaire distinct du CDT en fonction de la superficie dans chaque poststrate. Les effets aléatoires (nécessaires pour le calcul de l'EBLUP et de l'EDFM, mais non du prédicteur du M-quantile) ont été définis comme des effets régionaux indépendants (c'est-à-dire une spécification d'ordonnées à l'origine aléatoires) en se basant sur le fait que, dans l'échantillon original de l'AAGIS, la composante de variance entre régions explique environ 3 % de la variabilité résiduelle totale après suppression des deux valeurs aberrantes. Le but était d'estimer les moyennes régionales du CDT.

Les tableaux 5 et 6 montrent les biais relatifs médians et les REQM relatives médianes des divers estimateurs et des estimateurs correspondants de l'EQM de ces estimateurs en se basant sur les $K = 1\,000$ échantillons stratifiés indépendants tirés des populations albanaïenne et de l'AAGIS, respectivement. Il convient de souligner qu'en dépit du fait que les modèles linéaires mixtes ajustés aux données albanaises ainsi que de l'AAGIS semblent raisonnables, les gains dus à l'adoption de méthodes d'estimation sur petits domaines fondées sur ces modèles ne donnent pas lieu à des améliorations considérables de l'efficacité étant donné les tailles originales des échantillons régionaux pour ces enquêtes. Par ailleurs, l'estimateur du M-quantile, qui n'est pas fondé sur une spécification d'effets aléatoires, donne de bons résultats en ce qui concerne tant le biais que l'EQM pour la population de l'AAGIS dans ce cas (tableau 6, n_i médian = 53), tandis que l'EBLUP, même s'il donne les meilleurs résultats en ce qui concerne l'EQM pour la population albanaïenne (tableau 5, n_i médian = 56), affiche aussi les biais les plus élevés (mais néanmoins faibles, la valeur la plus grande étant inférieure à 2 %) pour les deux populations sachant la taille d'échantillon de domaine originale. L'estimateur par régression des données d'enquête donne de bons résultats, quoique, pour les deux

faible dans les domaines « aberrants ». Par contre, l'estimateur conditionnel de l'EQM pour l'EBLUP et pour l'EDFM suit d'assez près les REQM propres au domaine, tandis que le même estimateur de l'EQM fondé sur des poids M-quantile ont tendance à présenter un faible biais dans les domaines « ayant le comportement attendu » et un biais élevé dans les domaines « aberrants », ce qui représente, pourrait-on soutenir, un meilleur résultat que celui observé pour l'estimateur PR0 dans cette simulation. Il convient de souligner ici que, dans certaines circonstances, un modèle supposé peut être révisé après la détection de valeurs aberrantes. Cependant, cela nécessite un nombre suffisamment grand de valeurs aberrantes décelées pour permettre leur modélisation distincte, ce qui est peu probable en pratique. En outre, l'extrapolation de ces résultats au cas des très petites tailles d'échantillon de domaine doit se faire avec beaucoup de prudence, à cause de l'instabilité que peut manifester l'estimateur conditionnel de l'EQM dans ce cas.

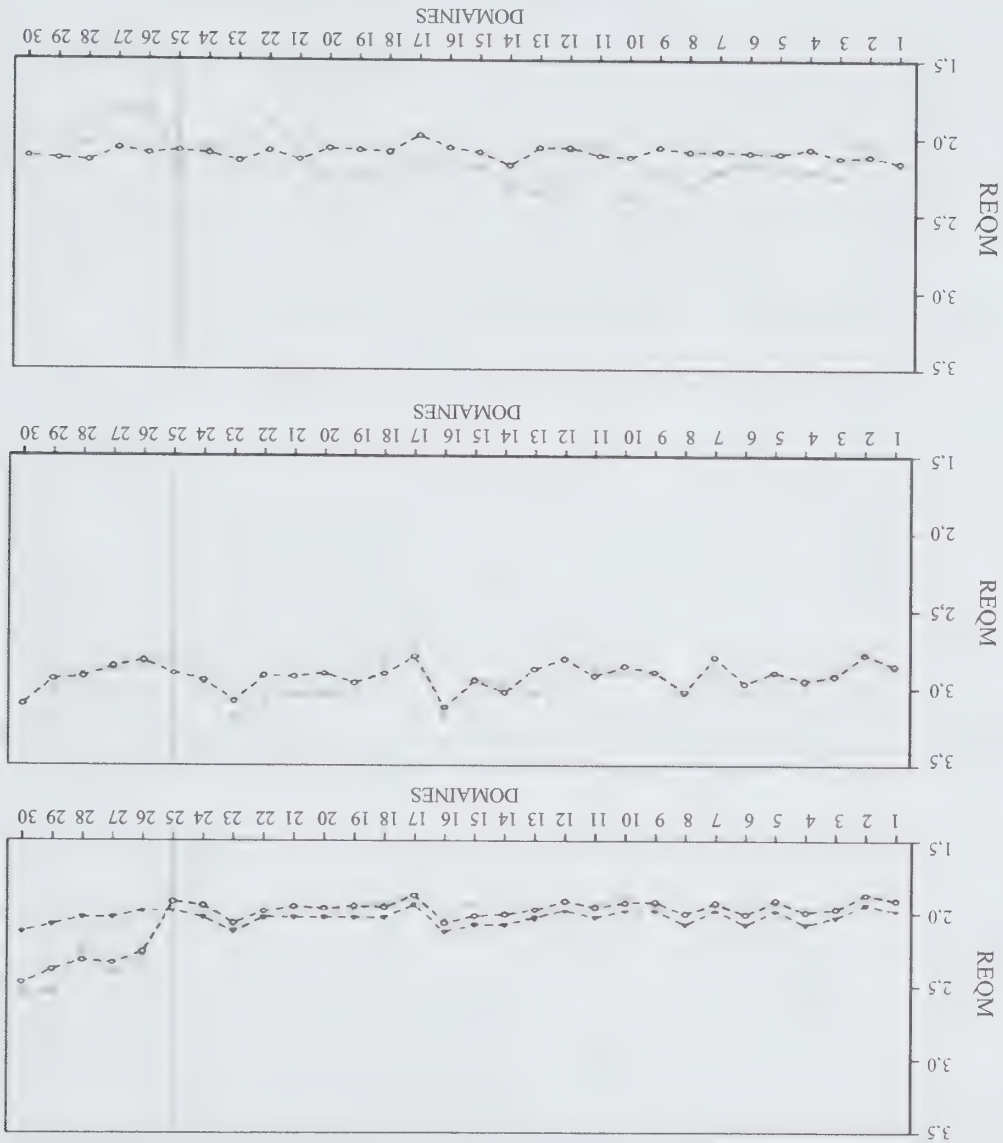


Figure 1 Valeurs propres au domaine de la REQM réelle (trait plein) et de la REQM estimée moyenne (trait interrompu) obtenues dans les simulations fondées sur un modèle de mélange SIM3-A et SIM3-A*. Les valeurs de l'estimateur PR0 sont indiquées par Δ , tandis que celles de l'estimateur conditionnel sont indiquées par \circ . Les graphiques montrent les résultats pour les estimateurs EBLUP (en haut), EDMF (au centre) et M-quantile (en bas). La droite verticale sépare les domaines 26 à 30 avec effets de « valeurs aberrantes » des domaines 1 à 25 se comportant « de la manière attendue ». La taille totale d'échantillon est de 600 avec des tailles d'échantillon de domaine égales à 20

Tableau 4
Biais relatif médian BR(M) et racine carrée de l'erreur quadratique moyenne relative médiane REQMR(M) pour les estimateurs de l'EQM dans les simulations fondées sur un modèle

Méthode de pondération	Estimateur de l'EQM	Simulation						
		SIM1-A	SIM1-B	SIM2-A	SIM2-B	SIM3-A	SIM3-B	SIM3-B*
Régression	VReg	7,59	21,82	11,81	20,78	23,66	34,27	23,97
	PR0	-0,83	-0,72	0,56	1,16	3,44	0,71	-15,65
	PR1	-0,97	-0,72	0,64	1,08	2,94	0,56	-13,70
	PR2	-0,92	-0,72	0,64	1,16	3,20	0,61	-14,65
	Conditionnel	3,89	-0,89	3,06	0,93	-0,05	-0,54	-2,56
EDFM, (18)	Conditionnel	-0,81	-0,80	-0,06	-0,42	-0,75	-0,75	-0,98
	Conditionnel	-3,10	-1,66	-0,09	-1,90	-5,04	-3,17	11,26
M-quantile, (22)	Conditionnel							11,04
REQMR(M), n_f médian = 20								
Régression	VReg	18	51	30	53	59	85	60
	PR0	12	7	15	10	11	7	29
	PR1	14	7	17	11	10	7	27
	PR2	12	7	16	10	11	7	28
	Conditionnel	62	31	70	49	31	30	42
EDFM, (18)	Conditionnel	70	70	126	128	71	71	67
	Conditionnel	32	34	49	48	31	32	48
M-quantile, (22)	Conditionnel							
BR(M), n_f médian = 5								
Régression	VReg	5,59	19,17	10,35	19,12	20,92	30,91	22,93
	PR0	3,51	-0,20	2,42	1,19	12,79	3,86	-30,64
	PR1	3,04	-0,50	2,13	1,00	10,84	3,10	-25,77
	PR2	3,16	-0,31	2,31	1,11	11,81	3,48	-28,16
	Conditionnel	37,52	4,38	24,11	8,93	8,18	1,50	-0,66
EDFM, (18)	Conditionnel	-0,24	-0,21	0,02	-0,09	-0,62	-0,33	1,29
	Conditionnel	-7,60	-6,17	5,70	5,00	-5,95	-5,60	5,89
M-quantile, (22)	Conditionnel							3,60
REQMR(M), n_f médian = 5								
Régression	VReg	17	46	33	51	54	78	59
	PR0	31	14	33	22	36	16	53
	PR1	48	18	44	28	34	16	48
	PR2	36	15	36	24	34	15	50
	Conditionnel	234	81	193	121	86	66	86
EDFM, (18)	Conditionnel	79	79	133	129	79	79	83
	Conditionnel	62	63	90	97	63	63	122
M-quantile, (22)	Conditionnel							102

L'estimateur conditionnel de l'EQM pour l'EBLUP présente un biais positif sous les scénarios normal (SIM1A) et du khi-carré (SIM2A), particulièrement dans le cas d'une corrélation intra-grappe modérée (3,89 % et 37,52 % pour la loi normale avec 20 et 5 unités dans chaque domaine respectivement, et 3,06 % et 24,11 % pour la loi du khi-carré avec 20 et 5 unités dans chaque domaine respectivement). Ce biais augmente lorsque la taille d'échantillon diminue. Cependant, la situation change quand nous examinons les résultats pour les composantes aberrantes des scénarios avec modèle de mélange (SIM3-A* et SIM3-B*). Ici, nous observons un biais négatif important pour les trois versions de l'estimateur PR (variant de -30,64 % à -5,81 % selon la taille d'échantillon de domaine). Comparativement, l'estimateur conditionnel de l'EQM pour l'EBLUP présente maintenant un biais négatif plus faible (-2,56 % et -0,66 %), tandis que le même estimateur de l'EQM appliqué à

L'estimateur du M-quantile présente un biais vers le haut. Pour l'EDFM, l'estimateur conditionnel de l'EQM est essentiellement sans biais. Comme, en ce qui concerne l'estimation de l'EQM, un biais positif est préférable à un biais négatif, il semble évident que l'estimateur conditionnel de l'EQM proposé permet de mieux traiter cette situation de valeur aberrante. La figure 1 illustre graphiquement ce point pour une taille d'échantillon de $n = 600$. Ici, nous montrons les REQMR propres au domaine et la moyenne (sur les simulations) des REQMR estimées dans chacun des 30 domaines pour les simulations de mélange SIM3-A et SIM3-A*, la droite verticale définissant les cinq domaines « aberrants ». Dans la partie supérieure de ce graphique, nous voyons que l'estimateur PR0 n'arrive pas à déceler la hausse de l'EQM de l'EBLUP pour ces domaines « aberrants », puisqu'ils présentent un biais légèrement élevé dans les domaines « se comportant de la manière attendue », puis un biais assez

Les résultats exposés au tableau 4 sont axés sur les biais médians $BR(M)$ et la racine carrée de l'erreur quadratique moyenne relative médiane $REQMR(M)$ des divers estimateurs de l'EQM. Naturellement, étant donné que toutes les hypothèses qui le sous-tendent sont satisfaites, l'estimateur $PR0$ et ses variantes propres au domaine, $PR1$ et $PR2$, donnent de très bons résultats dans les deux scénarios normaux ($SIM1-A$ et $SIM1-B$) et dans les deux scénarios du khi-carré ($SIM2-A$ et $SIM2-B$), le biais étant presque nul ($n_i = 20$) ou faible quand les tailles d'échantillon dans les domaines sont très faibles. Par contre, pour l'estimateur de l'EQM de l'estimateur par la régression synthétique, nous observons un biais relatif important sous tous les scénarios de simulation.

Tableau 1
Valeurs des paramètres utilisés dans les simulations fondées sur un modèle

Type	Simulation					
Gaussien	$SIM1-A$	10,40	40,32	94,09	0,1	$p = \sigma_2^2(\sigma_n^2 + \sigma_2^2)^{-1}$
khi-carré	$SIM1-B$	40,32	94,09	0,3	0,1667	
	$SIM2-A$	2,0	10,0	10,0	0,2857	
	$SIM2-B$	4,0	10,0	94,09	0,10	
Mélange (domaines 1 à 25)	$SIM3-A$	10,40	40,32	94,09	0,30	
	$SIM3-B$	225,0	94,09	0,7051		
Mélange (domaines 26 à 30)	$SIM3-A^*$	225,0	94,09	0,7051		
	$SIM3-B^*$	225,0	94,09	0,7051		

Tableau 2
Biais relatif médian $BR(m)$ et racine carrée de l'erreur quadratique moyenne relative médiane $REQMR(m)$ des estimateurs des moyennes de petit domaine dans les simulations fondées sur un modèle

Méthode de pondération		Simulation				
		$SIM1-A$	$SIM1-B$	$SIM2-A$	$SIM2-B$	$SIM3-A^*$
		Simulation				
		$BR(m), n_i$ médian = 20				
Régression	0,005	0,005	0,000	0,000	0,004	0,006
EBLUP, (13)	0,005	0,006	0,004	-0,002	0,004	0,006
EDFM, (18)	0,006	0,006	-0,005	-0,008	0,007	0,001
M-quantile, (22)	0,009	0,008	-0,002	0,002	0,015	-0,013
Régression	0,40	0,40	0,13	0,13	0,40	0,41
EBLUP, (13)	0,35	0,38	0,12	0,13	0,37	0,45
EDFM, (18)	0,55	0,55	0,41	0,43	0,56	0,55
M-quantile, (22)	0,41	0,41	0,13	0,13	0,41	0,36
		$BR(m), n_i$ médian = 5				
Régression	-0,002	-0,003	-0,001	0,002	-0,003	-0,004
EBLUP, (13)	0,001	0,005	-0,002	0,003	0,001	0,011
EDFM, (18)	-0,002	-0,002	-0,005	0,004	-0,001	-0,002
M-quantile, (22)	-0,001	-0,001	-0,001	-0,001	-0,003	-0,014
Régression	0,81	0,81	0,26	0,26	0,82	0,80
EBLUP, (13)	0,53	0,69	0,19	0,22	0,61	1,00
EDFM, (18)	1,13	1,13	0,83	0,83	1,13	1,13
M-quantile, (22)	0,81	0,81	0,26	0,26	0,81	0,80

Tableau 3
Définitions des estimateurs conditionnels de l'EQM pour diverses méthodes de pondération

Méthode de pondération	Définition de $\hat{\mu}_j, j \in I$		Estimateur de PQM
EBLUP (13)	(14)	(8)	(8)
EDFM (18)	(14)	(8)	(8)
M-quantile (22)	$x_j^T \hat{\beta}(\hat{q}_j)$	(7) avec $\lambda_i = 1$	(24) + (25)
EBLUP synthétique (23)	(14)		

domaines a été utilisée. Les tailles de population dans les petits domaines ont été réparties uniformément sur l'intervalle [443, 542] et maintenues fixes au cours des simulations. Dans chaque simulation, les valeurs de population de X ont été produites sous le modèle à ordonnées à l'origine aléatoires $y_j = 500 + 1,5x_j + u_j + e_j$, avec x_j tiré d'une loi du khi-carré avec 20 degrés de liberté. Les effets de domaine u_j et les effets individuels e_j ont été tirés indépendamment des lois $N(0, \sigma_u^2)$ et $N(0, \sigma_e^2)$ respectivement, avec les valeurs de σ_u et σ_e présentées aux lignes SIM1-A et SIM1-B du tableau 1. Un échantillon de taille $n = 600$ a été sélectionné dans chaque population simulée, avec des tailles d'échantillon de domaine proportionnelles aux populations de domaine fixes, ce qui a produit une taille médiane d'échantillon de domaine de $n_j = 20$. Le tirage des échantillons a été effectué par échantillonnage aléatoire stratifié, les strates étant définies par les petits domaines. Un total de $K = 1\,000$ simulations ont été exécutées.

Les conditions pour la deuxième étude en simulation fondée sur un modèle étaient les mêmes que pour la première, excepté que les effets aléatoires au niveau du domaine et les effets aléatoires au niveau de l'individu ont été tirés indépendamment de lois du khi-carré corrigées par la moyenne, respectivement. Les valeurs correspondantes des variances au niveau du domaine et au niveau individuel sont présentées aux lignes SIM2-A et SIM2-B dans le tableau 1. Enfin, dans la troisième étude en simulation fondée sur un modèle, les conditions ont été maintenues les mêmes que dans les simulations SIM1-A et SIM1-B pour les domaines 1 à 25, mais pour les domaines 26 à 30, les effets de domaine ont été tirés indépendamment d'une loi normale ayant une plus grande variance. Nous désignons ce cas comme un mélange dans le tableau 1, avec les variances pour les domaines 1 à 25 présentées aux lignes SIM3-A et SIM3-B, et les variances pour les domaines 26 à 30 présentées aux lignes SIM3-A* et SIM3-B*. Notre objectif dans cette troisième simulation était d'étudier le comportement des différentes méthodes d'estimation de l'EQM pour les domaines « aberrants » et nous présentons donc les valeurs pour les domaines 1 à 25 et pour les domaines 26 à 30 séparément dans les tableaux 2 et 4. Nous avons également reproduit chacun des trois scénarios susmentionnés en utilisant une plus petite taille globale d'échantillon de $n = 150$ (avec une taille médiane d'échantillon de domaine de $n_j = 5$). Ces simulations supplémentaires nous ont permis d'étudier l'effet de la réduction de la taille d'échantillon sur les propriétés des estimateurs de l'EQM.

Le tableau 2 montre le biais relatif médian $BR(m)$ et la racine carrée de l'erreur quadratique moyenne relative médiane $REQMR(m)$ des méthodes d'estimation sur petits domaines étudiées dans nos simulations pour les deux tailles d'échantillon ($n = 600$ et 150). Il s'agit de l'estimateur par

la régression synthétique (voir Rao 2003, page 136), de l'EBLUP avec les poids définis par (13), de l'EDFM avec les poids définis par (18) et de l'estimateur du M-quantile défini par les poids (22). Les différences entre les divers estimateurs sur petits domaines présentés dans le tableau 2 sont essentiellement celles attendues. Le biais n'est pas vraiment un problème (ce à quoi il fallait s'attendre, puisque les données de population suivent un modèle linéaire dans tous les cas), tandis que pour les scénarios de simulation 1 et 2, l'estimateur indirect (EBLUP) est le plus efficace si l'on s'en tient à la REQMR. L'estimateur du M-quantile est celui qui donne les meilleurs résultats pour SIM3-A* et SIM3-B* avec $n_j = 20$, mais son écart par rapport à l'estimateur par la régression synthétique se réduit pour le scénario avec les tailles d'échantillon de domaine plus petites. Notons que, dans ce cas, les poids du M-quantile (22) sont fondés sur une estimation robuste aux valeurs aberrantes du coefficient du M-quantile q_i pour le domaine i , défini par la médiane (plutôt que la moyenne) des coefficients des M-quantiles des unités échantillonnées dans ce domaine. En outre, à mesure que les tailles d'échantillon diminuent, la REQMR de tous les estimateurs augmente, mais la performance relative de ces estimateurs reste la même. Sous l'hypothèse de normalité, l'EBLUP est meilleur que l'estimateur du M-quantile, mais les différences entre ces deux estimateurs deviennent plus faibles à mesure que nous nous écartons de la normalité, tandis que l'estimateur du M-quantile est plus efficace dans les scénarios avec modèle de mélange.

Le tableau 3 présente les divers estimateurs de l'EQM étudiés dans nos simulations qui sont fondés sur l'approche proposée dans le présent article. Ils sont désignés collectivement comme des estimateurs « conditionnels » de l'EQM ci-après. Dans le tableau 4, nous montrons les propriétés des estimateurs de l'EQM pour les estimateurs sur petits domaines considérés au tableau 2. Notons qu'en plus des estimateurs conditionnels de l'EQM, nous fournissons les résultats pour trois autres estimateurs de l'EQM pour l'EBLUP, parmi lesquels PR0 désigne l'estimateur proposé par Prasad et Rao (1990), voir Rao (2003, section 6.2.6). Il convient de souligner que PR0 n'est pas un estimateur de l'EQM propre au domaine de l'EBLUP, mais de son EQM sous le modèle linéaire mixte (12), c'est-à-dire dont la moyenne est calculée sur les réalisations possibles de l'effet de domaine. Par contre, les estimateurs PR1 et PR2 de l'EQM présentés au tableau 4 sont les versions propres au domaine de l'estimateur PR0 proposé dans Rao (2003, section 6.3.2, expressions 6.3.15 et 6.3.16 respectivement). Enfin, nous notons que l'estimateur de l'EQM de l'estimateur par la régression synthétique que nous avons utilisé dans nos simulations est son estimateur de variance fondé sur un modèle de régression de population à effets fixes. Nous le désignons par VReg.

Il est clair que (23) est un estimateur pseudo-linéaire et nous pouvons donc utiliser (7) pour estimer sa variance de prédiction, en notant que, puisque $n_i = 0$, $a_{ij} = N_i w_{ij}^{\text{EBLUP}}$ et donc (7) devient

$$\hat{V}(m_i^{\text{SYN-EBLUP}}) = \sum_{j \in s_h} \{ (w_{ij}^{\text{SYN-EBLUP}})^2 + N_i^{-1} n_i^{-1} \{ \hat{\chi}_j^{-1} (y_j - \hat{\mu}_j) \}^2 \}. \quad (24)$$

Malheureusement, comme il n'y a pas d'échantillon dans le domaine i , nous ne pouvons pas utiliser (9) pour estimer le biais propre au domaine (2) de $m_i^{\text{SYN-EBLUP}}$. Cependant, sous le modèle mixte linéaire (12), la valeur prévue de ce biais est

$$E(m_i^{\text{SYN-EBLUP}} - m_i) =$$

$$\sum_{D^+} \sum_{j \in s_h} w_{ij}^{\text{SYN-EBLUP}} (x_j^T \beta + z_j^T u_h) - \bar{x}_i^T \beta - \bar{z}_i^T u_i.$$

L'espérance conditionnelle du carré de ce biais prévu, sachant les effets de domaine $u_s = (u_h; h = 1, \dots, D^+)$ pour les domaines échantillonnés, est

$$E\{E^2(m_i^{\text{SYN-EBLUP}} - m_i) | X, u_s\} =$$

$$\left\{ \sum_{D^+} \sum_{j \in s_h} w_{ij}^{\text{SYN-EBLUP}} (x_j^T \beta + z_j^T u_h) - \bar{x}_i^T \beta \right\}_2^2 + \bar{z}_i^T \Omega \bar{z}_i,$$

ce qui suggère immédiatement que, pour un domaine i non échantillonné, nous estimons le carré du biais de l'estimateur synthétique $m_i^{\text{SYN-EBLUP}}$ par

$$\hat{B}^2(m_i^{\text{SYN-EBLUP}}) = \left\{ \sum_{D^+} \sum_{j \in s_h} w_{ij}^{\text{SYN-EBLUP}} (x_j^T \beta^{\text{EBLUP}} + z_j^T \hat{u}_h) - \bar{x}_i^T \hat{\beta}^{\text{EBLUP}} \right\}_2^2 + \bar{z}_i^T \hat{\Omega} \bar{z}_i. \quad (25)$$

Ici, \hat{u}_h est l'effet estimé « détrencé » pour le domaine échantillonné h – voir (14). L'estimateur de l'EQM que nous proposons pour $m_i^{\text{SYN-EBLUP}}$ est alors la somme de (24) et (25). Notons que, contrairement à (8), cet estimateur de l'EQM ne contient aucune information provenant du domaine i , et n'est donc pas un estimateur de l'EQM propre au domaine de (23). En particulier, sa validité dépend entièrement du fait que le modèle mixte (12) est vérifié, et il n'est donc pas robuste à l'erreur de spécification de ce modèle.

3. Études en simulation de l'estimateur de l'EQM proposé

À la présente section, nous décrivons les résultats de cinq études en simulation destinées à évaluer la performance de l'approche d'estimation conditionnelle de l'EQM décrite à la section précédente. Trois de ces études sont des

dans chacune des cinq études, le principal indicateur de performance d'un estimateur de l'EQM est son biais relatif médian, défini par

$$\text{BR}(M) = \text{médiane} \left\{ M_i^{-1} K^{-1} \sum_{k=1}^K (\hat{M}_{ik} - M_i) \right\} \times 100.$$

Ici, l'indice inférieur i désigne les petits domaines et l'indice supérieur k , les K simulations Monte Carlo, avec \hat{M}_{ik} désignant la valeur de l'estimateur de l'EQM pour la simulation k dans le domaine i , et M_i désignant l'EQM réelle (c'est-à-dire Monte Carlo) dans le domaine i . Puisque nous préférons naturellement utiliser le plus stable de deux estimateurs de l'EQM approximativement sans biais, nous mesurons aussi la stabilité d'un estimateur de l'EQM par la médiane de la racine carrée de son erreur quadratique moyenne relative en pourcentage,

$$\text{REQMR}(M) = \text{médiane} \left\{ \left| K^{-1} \sum_{k=1}^K \left(\frac{\hat{M}_{ik} - M_i}{M_i} \right) \right|_2 \right\} \times 100.$$

Bien que le but du présent article ne soit pas de comparer diverses méthodes d'estimation sur petits domaines, il est utile de relier la performance de l'estimation de l'EQM pour une méthode particulière d'estimation sur petits domaines à la performance réelle d'estimation de cette méthode. Par conséquent, nous fournissons deux mesures de la performance relative des méthodes d'estimation sur petits domaines qui ont été utilisées dans nos simulations. Il s'agit de la médiane du biais relatif en pourcentage

$$\text{BR}(m) = \text{médiane} \left\{ m_i^{-1} K^{-1} \sum_{k=1}^K (\hat{m}_{ik} - m_{ik}) \right\} \times 100$$

et la médiane de la racine carrée de l'erreur quadratique moyenne relative en pourcentage

$$\text{REQMR}(m) = \text{médiane} \left\{ \left| K^{-1} \sum_{k=1}^K \left(\frac{\hat{m}_{ik} - m_{ik}}{m_{ik}} \right) \right|_2 \right\} \times 100$$

des estimations m_{ik} générées par une méthode d'estimation. Notons qu'ici, $\hat{m}_i = K^{-1} \sum_{k=1}^K m_{ik}$.

3.1 Simulations fondées sur un modèle

La première étude en simulation fondée sur un modèle s'appuyait sur des données de population générées sous le modèle mixte (12) avec des effets aléatoires gaussiens. Une taille de population de $N = 15\,000$, avec $D = 30$ petits

pour $h \neq i$, nous avons $\delta_i = 0$ pour l'EDFM et l'estimateur du biais (9) donne donc de bons résultats dans ce cas.

2.2.4 Estimation de l'EQM pour l'estimateur du M-quantile

Le troisième estimateur que nous envisageons est fondé sur l'approche de modélisation des M-quantiles décrite dans Chambers et Tzavidis (2006). Cette approche ne repose pas sur l'hypothèse d'un modèle mixte linéaire sous-jacent, s'appuyant plutôt sur la caractérisation de la relation entre y_j et x_j dans le domaine i en ce qui a trait au modèle de M-quantile linéaire qui est le mieux « adapté » aux valeurs d'échantillon y_j provenant de ce domaine. Autrement dit, cette approche consiste à remplacer (12) par un modèle de la forme

$$y_i = X_i' \beta(q_i) + e_i \quad (20)$$

où $\beta(q)$ désigne le vecteur de coefficients d'un modèle linéaire pour le M-quantile de régression d'ordre q pour les valeurs de population de X et q_i désigne le coefficient de M-quantile du domaine i . Étant donné une estimation \hat{q}_i de q_i , un algorithme des moindres carrés répétés itérativement (MCRl) est utilisé pour calculer une estimation

$$\hat{\beta}(q_i) = \{X_i' W_i(q_i) X_i\}^{-1} X_i' W_i(q_i) y_i \quad (21)$$

de $\beta(q_i)$ dans (20), et une valeur hors échantillon de y_j dans le domaine i est alors prédite par $\hat{y}_j = x_j' \hat{\beta}(q_i)$. Ici, $W_i(q_i)$ est la matrice diagonale des poids finaux utilisée dans l'algorithme MCRl.

Tzavidis, Marchetti et Chambers (2010) notent que la valeur de l'estimateur du M-quantile proposé dans Chambers et Tzavidis (2006) peut être interprétée comme la valeur prévue de Y dans le domaine i par rapport à un estimateur biaisé et la densité de probabilité de cette variable dans le domaine. Par conséquent, ils élaborent un estimateur amélioré du M-quantile, en remplaçant cet estimateur biaisé de la densité de probabilité par l'estimateur de la densité de Chambers et Dunstan (1986) sous le modèle propre au domaine (1). Cela revient à prédire m_i par

$$m_{\text{M}0}^i = \sum_{j \in s_n} w_{\text{M}0}^{ij} y_j = (w_{\text{M}0}^{ls})' y_s \quad (22)$$

où

$$w_{\text{M}0}^{ls} = n_{-1}^i \Delta_{-1}^{ls}$$

Ici, \bar{x}_{-1}^{ls} et \bar{x}_{-1}^{tr} sont les vecteurs des moyennes d'échantillon et hors échantillon des x_j dans le domaine i . Il n'est pas

difficile de montrer que les poids conformes à (22) sont calculés localement. En outre, si nous posons alors que $\hat{y}_j = x_j' \hat{\beta}(q_i)$, où $\hat{\beta}(q_i)$ est défini par (21), il est facile de voir que (9) est nul et que l'EQM propre au domaine de l'estimateur du M-quantile corrigé du biais (22) peut donc être estimé en utilisant simplement la composante de la variance de prédiction estimée (7). Puisque, dans (7), la constante λ_j est habituellement très proche de l'unité sous l'estimation du M-quantile, nous la fixons égale à cette valeur. Nous calculons les valeurs de (7) reliées à l'estimation sur petits domaines (EPD) sous l'approche de modélisation des M-quantiles.

Comme nous l'avons déjà fait pour l'EBLUP, nous notons que l'utilisation de (7) revient à traiter implicitement les poids définissant (22) comme fixes, ce qui n'est pas le cas en réalité, puisque la matrice $W_s(q_i)$ est une fonction des valeurs d'échantillon de X . Une conséquence directe est que la pseudo-linéarisation fondée sur l'estimation de l'EQM du prédicteur du M-quantile par la voie de (7) est une approximation de premier ordre de la valeur vraie de l'EQM de cet estimateur. Néanmoins, puisque tenir compte de la variabilité des poids dans la définition de l'estimateur du M-quantile complique considérablement l'estimation de son EQM – voir Street, Carroll et Ruppert (1988) pour un examen de cette question dans le contexte de la M-estimation « classique » des coefficients de régression – il est intéressant de voir comment l'estimateur (7) relativement simple se comporte quand il est utilisé pour estimer cette EQM.

2.3 L'estimation de l'EQM pour l'EBLUP synthétique pseudo-linéaire

Dans de nombreuses applications d'estimation sur petits domaines, il existe des domaines ne contenant pas d'unités échantillonnées et l'on utilise alors une estimation synthétique. Bien que ce genre d'estimateur n'entre pas dans la classe des estimateurs pseudo-linéaires examinés dans le présent article, les notions qui sous-tendent l'estimateur conditionnel de l'EQM (8) peuvent leur être appliquées également. Pour le voir, supposons que ces domaines sont numérotés les derniers, c'est-à-dire que si D^+ domaines ont un échantillon non nul, alors $n_h > 0$ pour $h \leq D^+$ et $n_h = 0$ pour $h > D^+$. Pour $i > D^+$ l'« EBLUP synthétique » pour m_i est

$$m_{\text{SYN-EBLUP}}^i = x_i' \hat{\beta}_{\text{EBLUP}} = (w_{\text{SYN-EBLUP}}^{ls})' y_s \quad (23)$$

$$\text{où} \quad \sum_{D^+} \sum_{h=1}^{n_h} w_{\text{SYN-EBLUP}}^{ij} y_j = (w_{\text{SYN-EBLUP}}^{ls})' x_i.$$

En supposant $n_i = m$, que $m\phi$ est petit et que N_i est grand, $\hat{M}_{\text{PR}}^i(m_{\text{EBLUP}}^i)$ est approximé par

$$\hat{M}_{\text{PR}}^i(m_{\text{EBLUP}}^i) \approx \hat{\sigma}_2^2 \{n_i^{-1} + 2(n_i - D)^{-1}\} + \hat{\sigma}_2^2. \quad (17)$$

En comparant (16) et (17), nous voyons que l'instabilité et la surestimation associées à l'utilisation de (8) dans cette situation sont toutes deux dues à l'emploi du carré du résidu au niveau du domaine à un seul degré de liberté $\bar{y}_{hs}^n - D^{-1} \sum_{h=1}^D \bar{y}_{hs}$ comme estimateur de σ_2^2 . Cela renforce les commentaires antérieurs voulant que (8) ne devrait généralement pas être utilisé pour estimer l'EQM de l'EBLUP si les tailles d'échantillon de domaine sont très petites ou, dans le cas particulier du modèle à moyennes aléatoires, si les tailles d'échantillon de domaine sont modérées quand la variabilité entre domaines est très faible comparativement à la variabilité à l'intérieur des domaines.

2.2.3 Estimation de l'EQM pour l'EDFM

Le deuxième prédicteur de m_i que nous considérons est l'estimateur direct fondé sur un modèle (EDFM) décrit dans Chandra et Chambers (2009). Cet estimateur est fondé sur le même modèle linéaire mixte (12) que l'EBLUP, le prédicteur EDFM étant défini comme

$$m_{\text{EDFM}}^i = \sum_{j \in s} w_{\text{EDFM}}^j y_j = (w_{\text{EDFM}}^T)^s y_s \quad (18)$$

où

$$w_{\text{EDFM}}^j = \frac{I(j \in s_i) w_{\text{EBLUP}}^j}{\sum_{k \in s} I(k \in s_i) w_{\text{EBLUP}}^k}. \quad (19)$$

Ici, $I(j \in s_i)$ est la fonction indiquant que l'unité j est dans l'échantillon du domaine i , et $w_{\text{EBLUP}}^j = (w_{\text{EBLUP}}^T)^j$ est le vecteur des poids qui définissent l'EBLUP pour le total de population y_j sous (12), c'est-à-dire

$$w_{\text{EBLUP}}^s = (w_{\text{EBLUP}}^T)^s = \mathbf{1}_n + \{H^s X^T + (\mathbf{1}_n - H^s X^T) \hat{\Sigma}^{-1} \hat{\Sigma}^{ss}\} \mathbf{1}_{N-n}$$

où $\mathbf{1}_n (\mathbf{1}_{N-n})$ désigne le vecteur unitaire de taille $n(N-n)$ et H^s a été défini à la section 2.2.1. Dans ce cas l'estimation par pseudo-linéarisation de l'EQM propre au domaine de l'EDFM est effectuée en se servant de (8), avec des poids définis par (19). Notons que les valeurs prévues estimées utilisées dans (8) lorsqu'on l'applique à l'EDFM sont les mêmes que les estimations détériorées (14) utilisées avec l'EBLUP, ce qui reflète le fait que l'EDFM et l'EBLUP sont tous deux fondés sur le même modèle linéaire mixte (12). Cependant, les poids pour l'EDFM (19) ne sont pas calculés localement, de sorte que le terme quadratique de biais dans (8) ne peut pas être ignoré quand on estime l'EQM de ce prédicteur. En outre, puisque

tandis que l'approximation correspondante de $\hat{V}(m_{\text{EBLUP}}^i)$ est

$$\hat{V}(\delta_i) = \hat{\sigma}_2^2 \sum_{h=1}^D \left(\frac{1 + m\phi}{1 + m\phi} \right)^2 \approx n^{-1} (1 + m\phi)^{-1} \hat{\sigma}_2^2,$$

il s'agit donc du terme principal dans cet estimateur dans cette situation. Cette expression peut être comparée à l'expression correspondante pour l'estimateur de l'EQM de l'EBLUP proposé par Prasad et Rao (1990). Sous le modèle à moyennes aléatoires, l'estimateur PR de l'EQM est

$$\hat{M}(m_{\text{EBLUP}}^i) \approx n^{-1} \left(D^{-1} \sum_{h=1}^D s_h^2 \right) + \left(\bar{y}_s - D^{-1} \sum_{h=1}^D \bar{y}_{hs} \right)^2. \quad (16)$$

Notons que l'espérance du résidu quadratique dans le second membre de (16) quand $m\phi$ est petit est $(D^{-1})(\sigma_n^2 + m^{-1}\sigma_2^2) = O(1)$, et il s'agit donc du terme principal dans cet estimateur dans cette situation. Cette expression peut être comparée à l'expression correspondante pour l'estimateur de l'EQM de l'EBLUP proposé par Prasad et Rao (1990). Sous le modèle à moyennes aléatoires, l'estimateur PR de l'EQM est

$$\hat{M}_{\text{PR}}(m_{\text{EBLUP}}^i) = (1 - f_i)^2 \hat{y}_i m_i \hat{\sigma}_2^2$$

$$+ (1 - \hat{y}_i)^2 \left(m \sum_{h=1}^D \hat{\tau}_h^{-1} + N_i^{-1} (1 - f_i)^{-1} \hat{\sigma}_2^2 \right) + \frac{T}{2} m \hat{\tau}_i^{-3} \left\{ \hat{\sigma}_4^4 \left(\frac{n - D}{n - D} \right) + \sum_{h=1}^D \hat{\tau}_h^{-2} \right\}$$

$$\text{où } \hat{\tau}_i = n_i \hat{\sigma}_2^2 + \hat{\sigma}_2^2 \text{ et } T = n - D \sum_{h=1}^D m_h^2 \hat{\tau}_h^{-2} + \left(\sum_{h=1}^D \hat{\tau}_h^{-2} \right) \left(\sum_{h=1}^D m_h^2 \hat{\tau}_h^{-2} \right) - \left(\sum_{h=1}^D m_h \hat{\tau}_h^{-2} \right)^2.$$

$$\hat{\mu}_i = \mathbf{x}_i^T \hat{\boldsymbol{\beta}}_{EBLUP} + (\bar{\mathbf{y}}_i - \bar{\mathbf{x}}_i^T \hat{\boldsymbol{\beta}}_{EBLUP})(\mathbf{x}_i - \bar{\mathbf{x}}_i)^T \hat{\boldsymbol{\beta}}_{EBLUP}$$

où $\bar{\mathbf{y}}_i$ et $\bar{\mathbf{x}}_i$ désignent les moyennes d'échantillon de Y et X respectivement, dans le domaine i . Sachant que (12) est le modèle de travail, une expression générale pour un tel estimateur « détecté » est

$$\hat{\mu}_i = \mathbf{x}_i^T \hat{\boldsymbol{\beta}}_{EBLUP} + \mathbf{z}_i^T \hat{\boldsymbol{\alpha}}_i \quad (14)$$

où $\hat{\boldsymbol{\alpha}}_i = (\mathbf{Z}^T \mathbf{Z}^{is})^{-1} \mathbf{Z}^T (\mathbf{y}^{is} - \mathbf{X}^{is} \hat{\boldsymbol{\beta}}_{EBLUP})$ est le prédicteur détecté de l'effet aléatoire pour le domaine i . Il n'est pas difficile de voir qu'alors, $\hat{\mu}_i = \sum_{k \in s} \phi_{kj} y_k$, où $\phi_{kj} = c_{ijsk} + b_{ijsk} I(k \in i)$, avec

$$c_{ijks} = (c_{ijks} : k \in s) = \hat{\Sigma}^{-1} \mathbf{X}^s (\mathbf{X}^s \hat{\Sigma}^{-1} \mathbf{X}^s - \mathbf{X}^{is} \mathbf{Z}^{is} (\mathbf{Z}^{is} \mathbf{Z}^{is})^{-1} \mathbf{Z}^{is})^T \mathbf{z}_j$$

et $\mathbf{b}^{ijks} = (b_{ijks} : k \in s) = \mathbf{Z}^{is} (\mathbf{Z}^{is} \mathbf{Z}^{is})^{-1} \mathbf{z}_j$. Notons que ces ϕ_{ij} sont également utilisés pour calculer la valeur de $\hat{\gamma}_j$ définie juste après (7).

Enfin, nous observons que, quand (14) est utilisé dans (8), le biais estimé (9) devient

$$B(\hat{\mu}_i) = \sum_D \left(\sum_{j \in s_h} \mathbf{w}_{EBLUP}^j \mathbf{z}_h^T \right) \mathbf{z}_h - \mathbf{z}_i^T \mathbf{u}_i$$

puisque les poids EBLUP (13) sont « calés localement » sur X , c'est-à-dire $\sum_{j \in s_h} \mathbf{w}_{EBLUP}^j \mathbf{x}_j = \bar{\mathbf{x}}_i$. Il s'ensuit que, dans ce cas, la variable δ_i définie juste avant (11) prend la forme

$$\delta_i = \sum_D \mathbf{w}_{EBLUP}^h \mathbf{z}_h^T \mathbf{u}_h - \mathbf{z}_i^T \mathbf{u}_i$$

où $\mathbf{w}_{EBLUP}^h = \sum_{j \in s_h} \mathbf{w}_{EBLUP}^j$. Pour une taille d'échantillon suffisamment grande, δ_i peut être approximé par

$$\delta_i \approx \sum_D \mathbf{w}_{EBLUP}^h \mathbf{z}_h^T (\mathbf{Z}^T \mathbf{Z}^{hs})^{-1} \mathbf{Z}^T (\mathbf{y}^{hs} - \mathbf{X}^{hs} \boldsymbol{\beta}) - \mathbf{z}_i^T (\mathbf{Z}^T \mathbf{Z}^{is})^{-1} \mathbf{Z}^T (\mathbf{y}^{is} - \mathbf{X}^{is} \boldsymbol{\beta})$$

où \mathbf{w}_{EBLUP}^h est l'équivalent BLUP de \mathbf{w}_{EBLUP}^h . La variance de δ_i peut par conséquent être estimée par

$$V(\delta_i) = \sum_D \mathbf{w}_{EBLUP}^h (\mathbf{Z}^T \mathbf{Z}^{hs})^{-1} \mathbf{z}_h \mathbf{z}_h^T \{\hat{\boldsymbol{\alpha}} + \hat{\boldsymbol{\sigma}}^2 (\mathbf{Z}^T \mathbf{Z}^{hs})^{-1} \mathbf{z}_h\} \quad (15)$$

Si $V(\delta_i)$ est petit comparativement à la valeur de (7) dans ce cas, on peut utiliser (8) pour estimer l'EOM de l'EBLUP. Cependant, si n_i est très petit, cette condition pourrait ne pas être vérifiée. Le cas échéant, il pourrait être souhaitable de considérer des estimateurs de l'EOM dépendant d'un modèle, tel que l'estimateur de l'EOM de Prasad et Rao (PR) (Prasad et Rao 1990 ; Rao 2003, section 7.2.3). Quand un modèle à moyennes aléatoires est utilisé,

2.2.2 Estimation de l'EOM pour l'EBLUP sous le

modèle à moyennes aléatoires

Le modèle à moyennes aléatoires est le cas particulier de (12) où $y_j = \beta + u_i + e_j$, avec $u_i \sim N(0, \sigma_u^2)$ et $e_j \sim N(0, \sigma^2)$. L'EBLUP de β est alors $\hat{\beta} = \sum_{h=1}^D \hat{\alpha}_h \bar{\mathbf{y}}_h$ avec $\hat{\alpha}_i = (\hat{\phi} + n_i^{-1})^{-1} \{ \sum_{h=1}^D (\hat{\phi} + n_h^{-1})^{-1} \}^{-1}$ et $\hat{\phi} = \hat{\sigma}_u^2 / \hat{\sigma}^2$, et l'EBLUP

(13) est défini par des poids de la forme

$$\mathbf{w}_{EBLUP}^j = (1 - f_j) (1 - \hat{\gamma}_i) \sum_{h=1}^D \hat{\alpha}_h n_h^{-1} I(j \in h) + \{ f_j + (1 - f_j) \hat{\gamma}_i \} n_i^{-1} I(j \in i)$$

avec $\hat{\gamma}_i = n_i \hat{\phi} (1 + n_i \hat{\phi})^{-1}$. Pour $j \in h$, $\hat{\mu}_j = \sum_{k \in s} \phi_{hj} y_k = \bar{\mathbf{y}}_h$ et donc

$$\hat{\gamma}_j = (1 - \phi_{jj})^2 + \sum_{k \in s(-j)} \phi_{kj}^2$$

$$= (1 - n_h^{-1})^2 + (n_h - 1) n_h^{-2} = (n_h - 1) n_h^{-1}$$

Il s'ensuit que l'estimateur (7) de la variance de prédiction conditionnelle de $\hat{\mu}_i$ est dans ce cas

$$V(\hat{\mu}_i)_{EBLUP} = (1 - f_i) \left[\sum_{h=1}^D \{ (1 - \hat{\gamma}_i) \}^2 \hat{\alpha}_h^2 n_h^{-2} \right]$$

où $s_h^2 = (n_h - 1)^{-1} \sum_{j \in s_h} (\mathbf{y}_j - \bar{\mathbf{y}}_h)^2$, tandis que selon (9), l'estimateur du biais conditionnel de prédiction de $\hat{\mu}_i$ est $B(\hat{\mu}_i)_{EBLUP} = (1 - f_i) (1 - \hat{\gamma}_i) (\hat{\beta} - \bar{\mathbf{y}}_i)$. Pour $h \neq i$, nous avons alors également

$$\mathbf{w}_{EBLUP}^h = \sum_{j \in s_h} \mathbf{w}_{EBLUP}^j$$

$$= (1 - f_i) \hat{\alpha}_h (1 + n_i \hat{\phi})^{-1} \approx \hat{\alpha}_h (1 + n_i \hat{\phi})^{-1}$$

quand nous ignorons les termes d'ordre $O(N_i^{-1})$. Une approximation semblable de (15) mène par conséquent à

$$V(\delta_i) = \sum_D \mathbf{w}_{EBLUP}^h (\mathbf{Z}^T \mathbf{Z}^{hs})^{-1} \mathbf{z}_h \mathbf{z}_h^T \{ \hat{\boldsymbol{\alpha}} + n_h^{-1} \hat{\sigma}^2 \}$$

$$\approx \hat{\sigma}^2 \sum_D \mathbf{w}_{EBLUP}^h \left(\frac{\hat{\alpha}_h}{1 + n_h \hat{\phi}} \right) \left(\frac{n_h}{1 + n_h \hat{\phi}} \right)$$

Supposons maintenant que la taille d'échantillon est la même dans chaque petit domaine, c'est-à-dire $n_i = m$. Alors, $n = mD$, $\hat{\alpha}_h = D^{-1}$ et l'approximation de $V(\delta_i)$ susmentionnée prend la forme

l'erreur de spécification des moments de deuxième ordre de X .

Un avantage important de l'expression (8) est qu'elle

peut être utilisée avec une gamme d'estimateurs sur petits domaines pouvant être exprimés sous forme pseudo-

linéaire. En particulier, de nombreux estimateurs sur petits

domaines élaborés sous des modèles qui sont des variantes

de (1) peuvent s'écrire sous cette forme, c'est-à-dire comme

des sommes pondérées des valeurs d'échantillon de X . Pour

illustrer ce point, nous nous concentrons maintenant sur

trois de ces estimateurs : l'EBLUP (Rao 2003, chapitre 6),

l'estimateur direct fondé sur un modèle (EDFM) de

Chandra et Chambers (2009) et le prédicteur du M-quantile

de Chambers et Tzavidis (2006). Chacun de ces estimateurs

peut s'écrire sous une forme pseudo-linéaire avec des poids

qui satisfont $w_j^* = O(n_j^{-1})$ pour $j \in s_1$ et $w_j^* = o(n_j^{-1})$

pour $j \notin s_1$, et donc (8) peut être utilisé.

2.2.1 Estimation de l'EQM pour l'EBLUP

Nous considérons d'abord l'EBLUP bien connu pour m_j

fondé sur l'extension du modèle (1) à un modèle mixte

linéaire au niveau de l'unité de la forme

$$y_j = X_j \beta + Z_j u_j + e_j \quad (12)$$

(12)

où y_j est le vecteur de dimension N_j des valeurs de popu-

lation de y_j dans le domaine i , X_j est la matrice de di-

mensions $N_j \times p$ correspondante des valeurs des variables

auxiliaires x_j , Z_j est la composante de dimensions $N_j \times q$

des X_j correspondant aux q composantes aléatoires de β ,

u_j est le vecteur de dimension q complexe d'effets aléatoires

propres au domaine et e_j est le vecteur de dimension N_j

des effets aléatoires individuels. On suppose habituellement

que les effets de domaine et les effets individuels sont

mutuellement indépendants, les effets de domaine étant

indépendants et identiquement distribués suivant la loi

$N(0, \Omega)$ et les effets individuels étant indépendants et

identiquement distribués suivant la loi $N(0, \sigma^2)$. Voir Rao

(2003, chapitre 6) pour le développement de la théorie qui

sous-tend ce prédicteur. Nous notons que l'EBLUP peut

s'écrire sous la forme pseudo-linéaire,

$$m_{j,EBLUP}^* = \sum_{j \in s} w_{j,EBLUP}^* y_j = (w_{j,EBLUP}^*)^T y_j \quad (13)$$

où

$$w_{j,EBLUP}^* = (w_{j,EBLUP}^*)^T$$

$$= N_j^{-1} [\Delta_{is} + \{H_T^s X_T^s + (I^n - H_T^s X_T^s) \Sigma_{ss}^{-1} \Sigma_{sp}^s\} \Delta_{ip}^s].$$

Ici, Δ_{ip}^s est un vecteur de taille $N - n$ qui « isole » les

unités non échantillonnées dans le domaine i , X_T^s et X_p^s

sont les matrices d'ordre $n \times p$ et $(N - n) \times p$ respective-

ment des valeurs dans l'échantillon et hors échantillon des

variables auxiliaires, I^n est la matrice identité d'ordre n ,

Etant donné ces conditions, l'estimation de l'EQM con-

ditionnelle de l'EBLUP peut être effectuée en se servant de

(8) avec les poids définis conformément à (13). À son tour,

cela requiert que nous ayons accès à des estimateurs sans

biais $\hat{\mu}_j$ des valeurs prévues individuelles propres au

domaine μ_j . Cependant, ces estimateurs peuvent être ins-

tables quand les tailles d'échantillon de domaine sont faibles.

Par conséquent, il est tentant de remplacer $\hat{\mu}_j$ par l'EBLUP

pour y_j , c'est-à-dire $y_{j,EBLUP}^* = x_j^T \beta_{j,EBLUP} + z_j^T u_{j,EBLUP}^*$, où

$\beta_{j,EBLUP}^*$ désigne le meilleur estimateur linéaire sans biais

empirique (EBLUE pour *Empirical Best Linear Unbiased*

Estimator) de β dans le modèle linéaire mixte (12) et

$u_{j,EBLUP}^*$ désigne l'effet de domaine prédit pour le domaine i

qui contient l'observation j . Malheureusement, en raison de

l'effet de rétrécissement bien connu associé aux EBLUP,

cette approche n'est pas recommandée. Pour l'illustrer, nous

notons que, dans (8), $V(\hat{m}_j)$ utilise $(y_j - \hat{\mu}_j)^2$ comme esti-

mateur de $E(y_j - \mu_j)^2$. Le biais dans cet estimateur est par

conséquent

$$E(y_j - \hat{\mu}_j)^2 - E(y_j - \mu_j)^2 = E(\hat{\mu}_j - \mu_j)^2 = -2E(y_j - \mu_j)(\hat{\mu}_j - \mu_j) + E(\hat{\mu}_j - \mu_j)^2 = -E\{\hat{\mu}_j - \mu_j\}(2y_j - \mu_j - \hat{\mu}_j) = -E\{\hat{\mu}_j - \mu_j\}$$

de sorte que nous attendons à ce que $V(\hat{m}_j)$ présente

un biais négatif si $E\{\hat{\mu}_j - \mu_j\}(2y_j - \mu_j - \hat{\mu}_j)$ est posi-

tif et inversement. Maintenant, soit l'unité d'échantillon j

provenant du domaine i et considérons le cas particulier

d'un modèle à ordonnées à l'origine aléatoires pour y_j , c'est-à-dire

$y_j = x_j^T \beta + u_j + e_j$, où u_j est l'effet aléatoire

pour le domaine i et e_j est un effet individuel aléatoire non

corrélé à u_j . Ici, $\mu_j = x_j^T \beta + u_j$. Supposons que nous

ayons une grande taille globale d'échantillon, ce qui nous

permet de remplacer $\beta_{j,EBLUP}^*$ par β . L'EBLUP $\hat{\mu}_j =$

$y_{j,EBLUP}^*$ peut alors être approximé par $\hat{\mu}_j = x_j^T \beta + y_j u_j$, où

y_j est un facteur de « rétrécissement ». Il s'ensuit que

$$(\hat{\mu}_j - \mu_j)(2y_j - \mu_j - \hat{\mu}_j) = 2u_j(y_j - 1)e_j - u_j^2(y_j - 1)^2$$

de sorte que $E(y_j - \hat{\mu}_j)^2 - E(y_j - \mu_j)^2 \approx (\gamma_i - 1)^2 \sigma_u^2$. Autre-

ment dit, nous nous attendons à ce que $V(\hat{m}_j)$ présente un

biais positif si nous utilisons l'EBLUP rétréci $y_{j,EBLUP}^*$ pour

définir $\hat{\mu}_j$. Nous notons aussi que ce biais disparaît (ap-

proximativement) si nous « dérétrécissons » la composante

résiduelle de cet EBLUP. Par exemple, dans le cas du

modèle à ordonnées à l'origine aléatoires très répandu, nous

utilisons

$$\hat{\sigma}^2 = \hat{\sigma}^2 = n^{-1} \sum_{j \in s} \left\{ (1 - \phi_{jj})^2 + \sum_{k \in s(-j)} \phi_{kj}^2 \right\}^{-1} (y_j - \hat{\mu}_j)^2.$$

Dans ce cas, (6) devient

$$V(\hat{m}_i) = N^{-2} \sum_{j \in s} \left\{ a_{ij}^2 + (N_i - n_i) n_i^{-1} \hat{\lambda}_i^{-1} (y_j - \hat{\mu}_j)^2 \right\}. \quad (7)$$

où, maintenant, $\hat{\lambda}_j = (1 - \phi_{jj})^2 + \sum_{k \in s(-j)} \phi_{kj}^2$. Comme toute

hypothèse concernant σ_j^2 dans l'extension de (1) à un

modèle de travail n'affecte que les termes de deuxième

ordre dans (3), l'estimateur (7) est robuste au biais, c'est-à-

dire qu'il demeure approximativement sans biais sous l'er-

reur de spécification des moments de deuxième ordre de ce

modèle de travail.

Un estimateur correspondant de l'EQM de m_i sous (1)

s'ensuit directement. Il s'agit de

$$M(\hat{m}_i) = V(\hat{m}_i) + B^2(\hat{m}_i), \quad (8)$$

où

$$B(m_i) = \sum_D \sum_{j \in s_h} w_{ij} \hat{u}_j - N_i^{-1} \sum_{j \in i} \hat{u}_j \quad (9)$$

est l'estimateur sans biais évident de (2).

L'utilisation du carré de l'estimateur sans biais (9) du

biais de \hat{m}_i dans l'estimateur conditionnel de l'EQM (8)

peut être critiquée, parce que ce terme n'est pas lui-même

sans biais pour le terme du carré du biais dans l'EQM. Cela

peut être corrigé en remplaçant (9) par

$$\hat{M}(\hat{m}_i) = V(\hat{m}_i) + B^2(\hat{m}_i) - V\{B(\hat{m}_i)\}, \quad (10)$$

où $V\{B(\hat{m}_i)\}$ est un estimateur approprié de la variance de

(9). Cependant, nous ne recommandons pas d'utiliser (10).

Pour voir pourquoi, soit $\hat{\beta} = D^{-1} \sum_{h=1}^D \beta_h$ et posons que

$\mathbf{d}_h = \hat{\beta}_h - \hat{\beta}$, où $\hat{\beta}_h$ est l'estimateur de β_h impliqué par les

poils ϕ_{kj} . En outre, posons que $w_{hi} = \sum_{j \in s_h} w_{ij}$ et $\bar{\mathbf{x}}^{whi} =$

$\sum_{j \in s_h} w_{hj} \bar{\mathbf{x}}_j$, de sorte que $\bar{\mathbf{x}}^{wi} = \sum_{j \in s} \sum_{h=1}^D w_{hj} \bar{\mathbf{x}}_j =$

$\sum_{h=1}^D w_{hi} \bar{\mathbf{x}}^{whi}$ est l'estimateur de $\bar{\mathbf{x}}$ fondé sur les poids w_{ij} .

Enfin, soit $\hat{\delta}_{hi} = \bar{\mathbf{x}}_i^T \mathbf{d}_h - \bar{\mathbf{x}}_i^T \mathbf{d}_i$ et posons que $\delta_i =$

$\sum_{h=1}^D w_{hi} \delta_{hi}$. Alors, (9) peut s'écrire

$$\hat{B}(\hat{m}_i) = (\bar{\mathbf{x}}^{wi} - \bar{\mathbf{x}}_i)^T \hat{\beta} + \left(\sum_D \sum_{h=1}^D w_{hi} \bar{\mathbf{x}}_i^T \mathbf{d}_h - \bar{\mathbf{x}}_i^T \mathbf{d}_i \right) = (\bar{\mathbf{x}}^{wi} - \bar{\mathbf{x}}_i)^T \hat{\beta} + \sum_D \sum_{h=1}^D w_{hi} (\bar{\mathbf{x}}^{whi} - \bar{\mathbf{x}}_h)^T \mathbf{d}_h - \bar{\mathbf{x}}_i^T \mathbf{d}_i \quad (11)$$

mis en balance avec la robustesse au biais de (8) sous

EQM. Cependant, cette sous-estimation éventuelle doit être

contribution de la variabilité de l'estimation de $\bar{\mathbf{T}}$ à cette

EQM vraie de l'EBLUP qui ne tient pas compte de la

variabilité fixes et (8) est par conséquent l'approximation de

l'EBLUP. Naturellement, les poids EBLUP ne sont pas

comme étant fixes et utiliser l'estimateur de l'EQM (8) pour

d'échantillon n , nous pouvons traiter les poids EBLUP

Autrement dit, pour de grandes valeurs de la taille globale

les poids EBLUP convergent vers les poids BLUP.

tion de $\bar{\mathbf{T}}$ converge vers la valeur vraie et, par conséquent,

modèle mixte linéaire est vraie, cet estimateur sur échan-

par substitution dans les poids BLUP. Si l'hypothèse du

mation du maximum de vraisemblance restreint ou MVR)

estimation sur échantillon efficace de $\bar{\mathbf{T}}$ (par exemple l'esti-

d'usage très répandu, est calculée en introduisant une

version empirique de ce prédicteur, c'est-à-dire l'EBLUP

dépendent de $\bar{\mathbf{T}}$ (voir Royall 1976). Par conséquent, la

somme pondérée des valeurs d'échantillon de Y où les poids

et la matrice de covariance $\bar{\mathbf{T}}$ peuvent s'écrire comme une

distribuées d'une variable aléatoire dont la valeur prévue β

β_i sont des réalisations indépendantes et identiquement

de (1) où les paramètres de régression propres au domaine

Predictor de m_i sous la variante du modèle mixte linéaire

linéaire sans biais (BLUP, pour *Best Linear Unbiased*

valeurs d'échantillon. Par exemple, le meilleur prédicteur

en structure, avec des pondérations qui dépendent de ces

pas cette condition, en ce sens qu'ils sont pseudo-linéaires

la plupart des estimateurs sur petits domaines ne satisfont

ne dépendent pas des valeurs d'échantillon de Y . Cependant,

selon laquelle les poids définissant l'estimateur linéaire \hat{m}_i

décrite à la sous-section précédente s'appuie sur l'hypothèse

L'approche de l'estimation conditionnelle de l'EQM

2.2 Estimation de l'EQM des estimateurs sur petits domaines pseudo-linéaires

domine (7), alors l'expression (8) ne devrait pas être utilisée.

2.2 Estimation de l'EQM des estimateurs sur petits domaines pseudo-linéaires

L'approche de l'estimation conditionnelle de l'EQM

décrite à la sous-section précédente s'appuie sur l'hypothèse

selon laquelle les poids définissant l'estimateur linéaire \hat{m}_i

ne dépendent pas des valeurs d'échantillon de Y . Cependant,

la plupart des estimateurs sur petits domaines ne satisfont

pas cette condition, en ce sens qu'ils sont pseudo-linéaires

en structure, avec des pondérations qui dépendent de ces

valeurs d'échantillon. Par exemple, le meilleur prédicteur

linéaire sans biais (BLUP, pour *Best Linear Unbiased*

Predictor) de m_i sous la variante du modèle mixte linéaire

de (1) où les paramètres de régression propres au domaine

β_i sont des réalisations indépendantes et identiquement

distribuées d'une variable aléatoire dont la valeur prévue β

et la matrice de covariance $\bar{\mathbf{T}}$ peuvent s'écrire comme une

somme pondérée des valeurs d'échantillon de Y où les poids

dépendent de $\bar{\mathbf{T}}$ (voir Royall 1976). Par conséquent, la

version empirique de ce prédicteur, c'est-à-dire l'EBLUP

d'usage très répandu, est calculée en introduisant une

estimation sur échantillon efficace de $\bar{\mathbf{T}}$ (par exemple l'esti-

mation du maximum de vraisemblance restreint ou MVR)

par substitution dans les poids BLUP. Si l'hypothèse du

modèle mixte linéaire est vraie, cet estimateur sur échan-

tion de $\bar{\mathbf{T}}$ converge vers la valeur vraie et, par conséquent,

les poids EBLUP convergent vers les poids BLUP.

Autrement dit, pour de grandes valeurs de la taille globale

d'échantillon n , nous pouvons traiter les poids EBLUP

comme étant fixes et utiliser l'estimateur de l'EQM (8) pour

l'EBLUP. Naturellement, les poids EBLUP ne sont pas

variabilité fixes et (8) est par conséquent l'approximation de

l'EQM vraie de l'EBLUP qui ne tient pas compte de la

contribution de la variabilité de l'estimation de $\bar{\mathbf{T}}$ à cette

EQM. Cependant, cette sous-estimation éventuelle doit être

mise en balance avec la robustesse au biais de (8) sous

EQM. Cependant, cette sous-estimation éventuelle doit être

contribution de la variabilité de l'estimation de $\bar{\mathbf{T}}$ à cette

EQM vraie de l'EBLUP qui ne tient pas compte de la

variabilité fixes et (8) est par conséquent l'approximation de

l'EBLUP. Naturellement, les poids EBLUP ne sont pas

comme étant fixes et utiliser l'estimateur de l'EQM (8) pour

d'échantillon n , nous pouvons traiter les poids EBLUP

Autrement dit, pour de grandes valeurs de la taille globale

les poids EBLUP convergent vers les poids BLUP.

tion de $\bar{\mathbf{T}}$ converge vers la valeur vraie et, par conséquent,

modèle mixte linéaire est vraie, cet estimateur sur échan-

par substitution dans les poids BLUP. Si l'hypothèse du

mation du maximum de vraisemblance restreint ou MVR)

estimation sur échantillon efficace de $\bar{\mathbf{T}}$ (par exemple l'esti-

d'usage très répandu, est calculée en introduisant une

version empirique de ce prédicteur, c'est-à-dire l'EBLUP

dépendent de $\bar{\mathbf{T}}$ (voir Royall 1976). Par conséquent, la

somme pondérée des valeurs d'échantillon de Y où les poids

et la matrice de covariance $\bar{\mathbf{T}}$ peuvent s'écrire comme une

distribuées d'une variable aléatoire dont la valeur prévue β

β_i sont des réalisations indépendantes et identiquement

de (1) où les paramètres de régression propres au domaine

Predictor de m_i sous la variante du modèle mixte linéaire

linéaire sans biais (BLUP, pour *Best Linear Unbiased*

valeurs d'échantillon. Par exemple, le meilleur prédicteur

en structure, avec des pondérations qui dépendent de ces

pas cette condition, en ce sens qu'ils sont pseudo-linéaires

la plupart des estimateurs sur petits domaines ne satisfont

ne dépendent pas des valeurs d'échantillon de Y . Cependant,

selon laquelle les poids définissant l'estimateur linéaire \hat{m}_i

décrite à la sous-section précédente s'appuie sur l'hypothèse

L'approche de l'estimation conditionnelle de l'EQM

décrite à la sous-section précédente s'appuie sur l'hypothèse

selon laquelle les poids définissant l'estimateur linéaire \hat{m}_i

ne dépendent pas des valeurs d'échantillon de Y . Cependant,

la plupart des estimateurs sur petits domaines ne satisfont

pas cette condition, en ce sens qu'ils sont pseudo-linéaires

en structure, avec des pondérations qui dépendent de ces

valeurs d'échantillon. Par exemple, le meilleur prédicteur

linéaire sans biais (BLUP, pour *Best Linear Unbiased*

Predictor) de m_i sous la variante du modèle mixte linéaire

de (1) où les paramètres de régression propres au domaine

β_i sont des réalisations indépendantes et identiquement

distribuées d'une variable aléatoire dont la valeur prévue β

et la matrice de covariance $\bar{\mathbf{T}}$ peuvent s'écrire comme une

somme pondérée des valeurs d'échantillon de Y où les poids

dépendent de $\bar{\mathbf{T}}$ (voir Royall 1976). Par conséquent, la

version empirique de ce prédicteur, c'est-à-dire l'EBLUP

d'usage très répandu, est calculée en introduisant une

estimation sur échantillon efficace de $\bar{\mathbf{T}}$ (par exemple l'esti-

mation du maximum de vraisemblance restreint ou MVR)

par substitution dans les poids BLUP. Si l'hypothèse du

modèle mixte linéaire est vraie, cet estimateur sur échan-

tion de $\bar{\mathbf{T}}$ converge vers la valeur vraie et, par conséquent,

les poids EBLUP convergent vers les poids BLUP.

Autrement dit, pour de grandes valeurs de la taille globale

d'échantillon n , nous pouvons traiter les poids EBLUP

comme étant fixes et utiliser l'estimateur de l'EQM (8) pour

l'EBLUP. Naturellement, les poids EBLUP ne sont pas

variabilité fixes et (8) est par conséquent l'approximation de

l'EQM vraie de l'EBLUP qui ne tient pas compte de la

contribution de la variabilité de l'estimation de $\bar{\mathbf{T}}$ à cette

EQM. Cependant, cette sous-estimation éventuelle doit être

mise en balance avec la robustesse au biais de (8) sous

EQM. Cependant, cette sous-estimation éventuelle doit être

contribution de la variabilité de l'estimation de $\bar{\mathbf{T}}$ à cette

EQM vraie de l'EBLUP qui ne tient pas compte de la

variabilité fixes et (8) est par conséquent l'approximation de

l'EBLUP. Naturellement, les poids EBLUP ne sont pas

comme étant fixes et utiliser l'estimateur de l'EQM (8) pour

d'échantillon n , nous pouvons traiter les poids EBLUP

Autrement dit, pour de grandes valeurs de la taille globale

les poids EBLUP convergent vers les poids BLUP.

tion de $\bar{\mathbf{T}}$ converge vers la valeur vraie et, par conséquent,

modèle mixte linéaire est vraie, cet estimateur sur échan-

par substitution dans les poids BLUP. Si l'hypothèse du

mation du maximum de vraisemblance restreint ou MVR)

estimation sur échantillon efficace de $\bar{\mathbf{T}}$ (par exemple l'esti-

d'usage très répandu, est calculée en introduisant une

version empirique de ce prédicteur, c'est-à-dire l'EBLUP

dépendent de $\bar{\mathbf{T}}$ (voir Royall 1976). Par conséquent, la

somme pondérée des valeurs d'échantillon de Y où les poids

et la matrice de covariance $\bar{\mathbf{T}}$ peuvent s'écrire comme une

distribuées d'une variable aléatoire dont la valeur prévue β

β_i sont des réalisations indépendantes et identiquement

de (1) où les paramètres de régression propres au domaine

Predictor de m_i sous la variante du modèle mixte linéaire

linéaire sans biais (BLUP, pour *Best Linear Unbiased*

valeurs d'échantillon. Par exemple, le meilleur prédicteur

en structure, avec des pondérations qui dépendent de ces

pas cette condition, en ce sens qu'ils sont pseudo-linéaires

la plupart des estimateurs sur petits domaines ne satisfont

ne dépendent pas des valeurs d'échantillon de Y . Cependant,

selon laquelle les poids définissant l'estimateur linéaire \hat{m}_i

décrite à la sous-section précédente s'appuie sur l'hypothèse

L'approche de l'estimation conditionnelle de l'EQM

décrite à la sous-section précédente s'appuie sur l'hypothèse

selon laquelle les poids définissant l'estimateur linéaire \hat{m}_i

ne dépendent pas des valeurs d'échantillon de Y . Cependant,

la plupart des estimateurs sur petits domaines ne satisfont

pas cette condition, en ce sens qu'ils sont pseudo-linéaires

en structure, avec des pondérations qui dépendent de ces

valeurs d'échantillon. Par exemple, le meilleur prédicteur

linéaire sans biais (BLUP, pour *Best Linear Unbiased*

Predictor) de m_i sous la variante du modèle mixte linéaire

de (1) où les paramètres de régression propres au domaine

β_i sont des réalisations indépendantes et identiquement

distribuées d'une variable aléatoire dont la valeur prévue β

et la matrice de covariance $\bar{\mathbf{T}}$ peuvent s'écrire comme une

somme pondérée des valeurs d'échantillon de Y où les poids

dépendent de $\bar{\mathbf{T}}$ (voir Royall 1976). Par conséquent, la

version empirique de ce prédicteur, c'est-à-dire l'EBLUP

d'usage très répandu, est calculée en introduisant une

estimation sur échantillon efficace de $\bar{\mathbf{T}}$ (par exemple l'esti-

mation du maximum de vraisemblance restreint ou MVR)

par substitution dans les poids BLUP. Si l'hypothèse du

modèle mixte linéaire est vraie, cet estimateur sur échan-

tion de $\bar{\mathbf{T}}$ converge vers la valeur vraie et, par conséquent,

les poids EBLUP convergent vers les poids BLUP.

Autrement dit, pour de grandes valeurs de la taille globale

d'échantillon n , nous pouvons traiter les poids EBLUP

comme étant fixes et utiliser l'estimateur de l'EQM (8) pour

l'EBLUP. Naturellement, les poids EBLUP ne sont pas

que ces poids ne dépendent pas des valeurs d'échantillon de X . En outre, nous supposons que $w_{ij} = O(n_i^{-1})$ pour $j \in s_i$, $w_{ij} = o(n_i^{-1})$ pour $j \notin s_i$, et $\sum_{j \in s_i} w_{ij} = 1$. Ici, s_i désigne les n_i unités d'échantillon provenant du domaine i . Le biais de \hat{m}_i sous (1) est alors

$$E(\hat{m}_i - m_i) = \left(\sum_{h=1}^D \sum_{j \in s_h} w_{ij} x_j^T \beta_h \right) - \bar{x}_i^T \beta_i, \quad (2)$$

où \bar{x}_i désigne le vecteur de valeurs moyennes des variables auxiliaires dans le domaine i . De même, la variance de prédiction de \hat{m}_i sous (1) est

$$\text{Var}(\hat{m}_i - m_i) = N_i^{-2} \left\{ \sum_{h=1}^D \sum_{j \in s_h} a_{ij}^2 \sigma_j^2 + \sum_{j \in \pi_i} \sigma_j^2 \right\}, \quad (3)$$

où r_i désigne les unités non échantillonnées dans le domaine i et $a_{ij} = N_i w_{ij} - I(j \in i)$. Nous utilisons $I(A)$ pour désigner la fonction indicatrice de l'événement A , de sorte que $I(j \in i)$ prend la valeur 1 si l'unité de population j provient du domaine i et la valeur zéro autrement. Notons que, puisque a_{ij} est d'ordre $O(N_i n_i^{-1})$ pour $j \in s_i$, le premier terme entre parenthèses dans (3) est le terme principal de cette variance de prédiction si N_i est grand comparativement à n_i .

Soit $j \in h$. Nous considérons le cas particulier important où $\mu_j = E(y_j | x_j) = x_j^T \beta_h$ est estimé par $\hat{\mu}_j = x_j^T \hat{\beta}_h = \sum_{k \in s} \phi_{hj} y_k$ avec les ϕ_{hj} correspondant aux poids appropriés. Alors

$$y_j - \hat{\mu}_j = (1 - \phi_{jj}) y_j - \sum_{k \in s(-j)} \phi_{kj} y_k$$

et donc

$$\text{Var}(y_j - \hat{\mu}_j) = \sigma_j^2 \left\{ (1 - \phi_{jj})^2 + \sum_{k \in s(-j)} \phi_{kj}^2 (\sigma_k^2 / \sigma_j^2) \right\} \quad (4)$$

sous (1). Ici, $s(-j)$ désigne l'échantillon s dont l'unité j est exclue. Si, en outre, $\hat{\mu}_j$ est sans biais pour μ_j sous (1), c'est-à-dire

$$E(y_j - \hat{\mu}_j) = 0, \quad (5)$$

nous pouvons adopter l'approche de Royall et Cumberland (1978) et estimer (3) par

$$\hat{V}(\hat{m}_i) = N_i^{-2} \left\{ \sum_{h=1}^D \sum_{j \in s_h} a_{ij}^2 \hat{\lambda}_{j-1} (y_j - \hat{\mu}_j)^2 + \sum_{j \in \pi_i} \hat{\sigma}_j^2 \right\}, \quad (6)$$

où $\hat{\lambda}_j = (1 - \phi_{jj})^2 + \sum_{k \in s(-j)} \hat{\gamma}_{kj}^2 \phi_{kj}^2$ et $\hat{\gamma}_{hj} = \hat{\sigma}_k^2 / \hat{\sigma}_j^2$. Habituellement, les estimations $\hat{\sigma}_j^2$ des variances résiduelles dans (6) sont calculées sous un perfectionnement de « modèle de travail » de (1). Dans la situation qui nous concerne le plus, où les tailles d'échantillon dans les différents domaines sont trop faibles pour estimer fiablement la variabilité propre au domaine, une hypothèse de groupement peut être faite, c'est-à-dire $\sigma_j^2 = \sigma^2$, auquel cas nous posons

domaines non chevauchants, qui contiennent chacun des unités échantillonnées, les tailles d'échantillon réalisées dans chacun des domaines échantillonnés étant faibles. Nous supposons aussi qu'il existe un nombre connu N_i d'unités de la population dans le domaine i et qu'un nombre n_i de ces unités sont échantillonnées. Le nombre total d'unités dans la population est $N = \sum_{i=1}^D N_i$, et la taille d'échantillon totale correspondante est $n = \sum_{i=1}^D n_i$. Dans la suite, nous utilisons s pour désigner l'ensemble d'unités dans l'échantillon, s_i étant le sous-ensemble tiré du domaine i , et nous utilisons des expressions telles que $j \in i$ et $j \in s$ pour faire référence aux unités qui constituent le domaine i et l'échantillon s , respectivement.

Des modèles linéaires sont souvent utilisés pour fonder les estimateurs des moyennes de population. Cependant, quand sont requises des estimations des moyennes de domaines correspondantes, il n'est habituellement pas raisonnable de supposer qu'un modèle linéaire applicable à l'ensemble de la population s'applique aussi dans chaque domaine. Par conséquent, nous adoptons une approche conditionnelle et considérons l'estimation de l'EQM des estimateurs des moyennes de domaine quand les modèles linéaires à appliquer ne sont pas les mêmes dans les différents domaines. En particulier, nous nous penchons sur les estimateurs qui peuvent être exprimés comme une somme pondérée des valeurs d'échantillon, en les qualifiant de « linéaire » dans la suite pour indiquer qu'ils ont une structure linéaire.

Pour commencer, soit y_j la valeur de Y pour l'unité j de la population et supposons que cette unité est dans le domaine i . Nous émettons aussi l'hypothèse d'un modèle linéaire propre au domaine pour y_j de la forme

$$y_j = x_j^T \beta_i + e_j. \quad (1)$$

Ici, x_j est un vecteur de dimension $p \times 1$ de variables auxiliaires au niveau de l'unité pour l'unité j , β_i est un vecteur de dimension $p \times 1$ de coefficients de régression propres au domaine et e_j est un effet aléatoire au niveau de l'unité de moyenne nulle et de variance σ_j^2 pour lequel il n'existe pas de corrélation entre les diverses unités de population. Nous ne faisons aucune hypothèse au sujet de σ_j^2 à ce stade. Notons que, tout au long de l'exposé, nous supposons que la méthode d'échantillonnage utilisée est non informative pour les valeurs de population de Y sachant les valeurs correspondantes des variables auxiliaires et les appartenances aux domaines des unités de population. Par conséquent, (1) s'applique à la fois au niveau de l'échantillon et de la population.

Soit y_s le vecteur colonne de valeurs d'échantillon de Y_j et soit $w_{is}^T = \{w_{ij}; j \in s\}$ le vecteur colonne de poids fixes tels que $\hat{m}_i = w_{is}^T y_s = \sum_{j \in s} w_{ij} y_j$ est un estimateur linéaire de $m_i = N_i^{-1} \sum_{j \in i} y_j$. Par « fixes » nous entendons ici

Estimation de l'erreur quadratique moyenne robuste au biais pour les estimateurs sur petits domaines pseudo linéaires

Ray Chambers, Hukum Chandra et Nikos Tzavidis¹

Résumé

Nous proposons une méthode d'estimation de l'erreur quadratique moyenne (EQM) pour les estimateurs des moyennes de domaine en population finie qui peuvent être exprimés sous une forme pseudo-linéaire, c'est-à-dire comme une somme pondérée des valeurs d'échantillon. En particulier, la méthode proposée peut être utilisée pour estimer l'EQM du meilleur prédicteur linéaire sans biais empirique, de l'estimateur direct fondé sur un modèle et du prédicteur du M-quantile. Elle représente une extension des idées de Royall et Cumberland (1978) et mène à des estimateurs de l'EQM qui sont plus simples à mettre en œuvre et éventuellement plus robustes au biais que ceux proposés dans la littérature sur les petits domaines. Cependant, il convient de souligner que les estimateurs de l'EQM définis en utilisant cette méthode peuvent également présenter une grande variabilité quand les tailles d'échantillon de domaine sont très petites. Nous illustrons les propriétés de la méthode à l'aide de simulations à grande échelle sous un modèle et sous un plan de sondage, dans ce dernier cas en nous fondant sur deux ensembles de données d'enquête réels contenant des données sur des petits domaines.

Mots clés : Meilleur prédicteur linéaire sans biais ; modèle du M-quantile ; estimation directe fondée sur un modèle ; modèle à effets aléatoires ; estimation sur petits domaines.

1. Introduction

Les modèles linéaires, et les prédicteurs linéaires fondés sur ces modèles, sont d'usage très répandu pour l'inférence d'après des données d'enquête. Cependant, ces modèles risquent d'être spécifiés incorrectement, particulièrement en ce qui concerne les moments de deuxième ordre et d'ordre plus élevé. Des méthodes d'estimation de l'erreur quadratique moyenne (EQM) des prédicteurs linéaires de quantités de population finie robustes au biais, c'est-à-dire des méthodes qui demeurent approximativement sans biais lorsque les hypothèses au sujet des moments de deuxième ordre et d'ordre plus élevé sont violées, ont été élaborées. Valliant, Dorfman et Royall (2000, chapitre 5) discutent de l'estimation de l'EQM robuste au biais pour de tels prédicteurs quand on suppose qu'une population suit un modèle linéaire. Dans le présent article, nous abordons un problème subsidiaire, qui est celui de l'estimation de l'EQM robuste au biais des estimateurs des moyennes de domaine de population finie qui peuvent être exprimés sous une forme pseudo-linéaire, c'est-à-dire comme des sommes pondérées, mais dans lesquels les poids peuvent dépendre des valeurs d'échantillon de la variable d'intérêt. Une application importante, et qui motive notre approche, est l'inférence sur petits domaines. Par conséquent, dans la suite de l'exposé, le petit domaine sera le domaine d'intérêt. Notre approche, qui représente une extension des idées de Royall et Cumberland (1978), semble produire des estimateurs de l'EQM plus simples à mettre en œuvre que ceux qui ont été proposés dans la littérature sur les petits domaines.

Le plan de l'article est le suivant. À la section 2, nous discutons de l'estimation de l'EQM sous un modèle linéaire propre au domaine. Autrement dit, nous nous concentrons sur l'estimation de l'EQM conditionnelle. Puis, nous montrons comment notre approche pourrait être utilisée pour estimer l'EQM de trois prédicteurs linéaires sur petits domaines différents quand ils sont exprimés sous une forme pseudo-linéaire, à savoir a) le meilleur prédicteur linéaire sans biais empirique ou EBLUP (pour *empirical best linear unbiased predictor*) (Henderson 1953), b) l'estimateur direct fondé sur un modèle (EDFM) de Chandra et Chambers (2009) et c) le prédicteur du M-quantile (Chambers et Tzavidis 2006). À la section 3, nous présentons les résultats d'une série d'études en simulation qui illustrent les propriétés sous le modèle et sous le plan de notre approche d'estimation de l'EQM. Enfin, à la section 4, nous résumons nos principaux résultats. Tout au long de l'exposé, nous utilisons i ou h comme indice des D petits domaines d'intérêt, et j ou k comme indice des unités de population distinctes dans ces domaines.

2. Estimation de l'EQM robuste au biais pour les estimateurs pseudo-linéaires

2.1 Estimation de l'EQM sous un modèle linéaire propre au domaine

Nous considérons la situation où nous avons une population finie de taille N dont est tiré un échantillon de taille n . Nous supposons que cette population est constituée de D

1. Ray Chambers, Centre for Statistical and Survey Methodology, University of Wollongong, Wollongong, NSW, 2522, Australie. Courriel : raychandra@uow.edu.au ; Hukum Chandra, Indian Agricultural Statistics Research Institute, Library Avenue, New Delhi-110012, Inde. Courriel : hchandra@iasri.res.in ; Nikos Tzavidis, Social Statistics and Southampton Statistical Sciences Research Institute, University of Southampton, Southampton, SO17 1BJ, R.-U. Courriel : n.tzavidis@soton.ac.uk.

$$A^{-1} \geq \frac{(d-c)^{g+h-2} B(a+g-1, b+h-1)}{B(a, b) B(g, h) \{F_{g,h}^g(d) - F_{g,h}^g(c)\}} = H_1 > 0.$$

En outre, nous avons

$$A^{-1} \leq \int_d^c f_1(x) \sup_{c < x < d} f_2(x) dx.$$

Donc, en vertu du lemme 2 (a),

$$A^{-1} \leq \frac{\delta^{a-1} (1-\delta)^{b-1}}{\delta^{a-1} (1-\delta)^{b-1}} \int_d^c f_1(x) B(a, b) (d-c) B(a, b) dx = \frac{(d-c) B(a, b)}{\delta^{a-1} (1-\delta)^{b-1}} = H_2 > 0.$$

Preuve du théorème 2

Pour prouver cette allégation, nous calculons la fonction de répartition $F^X(\cdot)$ de la variable aléatoire X définie dans le théorème. Nous avons

$$F^X(x) = P(X \leq x)$$

$$= P[F^{-1}\{UF_{g,h}^g(d) + (1-U)F_{g,h}^g(c)\} \leq x]$$

$$= P[UF_{g,h}^g(d) + (1-U)F_{g,h}^g(c) \leq F_{g,h}^g(x)]$$

$$= P\{U\{F_{g,h}^g(d) - F_{g,h}^g(c)\} \leq F_{g,h}^g(x) - F_{g,h}^g(c)\}$$

$$= P\left[U \leq \frac{F_{g,h}^g(x) - F_{g,h}^g(c)}{F_{g,h}^g(d) - F_{g,h}^g(c)}\right].$$

Maintenant, puisque $U \sim \text{Uniforme}(0, 1)$, d'après l'expression susmentionnée pour $F^X(\cdot)$, nous avons $F^X(x) = 1$ si $x \geq d$ et $F^X(x) = 0$ si $x \leq c$. Quand $c \leq x \leq d$, nous avons

$$F^X(x) = \frac{F_{g,h}^g(x) - F_{g,h}^g(c)}{F_{g,h}^g(d) - F_{g,h}^g(c)}.$$

Cela montre que X a la densité bêta tronquée $f_1(x)$ par (20).

Maintenant, si nous voulons utiliser l'algorithme d'acceptation-rejet, considérons

$$\frac{f(x)}{Af_z(x)} = \frac{f_1(x)}{Af_z(x)}.$$

En vertu du lemme 2, nous avons

$$\sup_{c < \pi < d} \left\{ \frac{f(x)}{f_1(x)} \right\} = A \sup_{c < \pi < d} f_2(x) = A \frac{\delta^{a-1} (1-\delta)^{b-1}}{(d-c) B(a, b)} > \infty.$$

Donc, en vertu de l'algorithme d'acceptation-rejet, si

$$A \leq \frac{1}{(d-c)^{a+b-2}} \left(\frac{\delta}{X-c} \right)^{a-1} \left(\frac{d-X}{1-\delta} \right)^{b-1},$$

alors X a la densité $f(x)$ par (19).

Bibliographie

- Cochran, W.G. (1977). *Sampling Techniques*, troisième édition. New York : John Wiley & Sons, Inc.
- Gilks, W.R., et Wild, P. (1992). Adaptive rejection sampling for gibbs sampling. *Journal of the Royal Statistical Society, Séries C*, 41, 337-348.
- Gelman, A. (2006). Prior distribution for variance parameters in hierarchical models. *Bayesian Analysis*, 1, 515-533.
- Ghosh, M., Natarajan, K., Stroud, T.W.F., et Carlin, B.P. (1998). Generalized linear models for small-area estimation. *Journal of the American Statistical Association*, 93, 273-282.
- Hillmer, S.C., et Trabelsi, A. (1987). Benchmarking of economic time series. *Journal of the American Statistical Association*, 82, 1064-1071.
- Lazar, R., Meeden, G. et Nelson, D. (2008). Une approche bayésienne non informative de l'échantillonnage d'une population finie en utilisant des variables auxiliaires. *Techniques d'enquête*, 34, 55-70.
- Nandram, B. (1998). A Bayesian analysis of the three-stage hierarchical multinomial model. *Journal of Statistical Computation and Simulation*, 61, 97-126.
- Nandram, B., et Choi, J.W. (2002). Hierarchical Bayesian nonresponse models for binary data from small areas with uncertainty about ignorability. *Journal of the American Statistical Association*, 97, 381-388.
- Nandram, B., et Choi, J.W. (2002). A Bayesian analysis of a proportion under non-ignorable non-response. *Statistics in Medicine*, 21, 9, 1189-1212.
- Nandram, B., et Sedransk, J. (1993). Bayesian predictive inference for a finite population proportion: Two-stage cluster sampling. *Journal of the Royal Statistical Society, Séries B*, 55, 399-408.
- Nandram, B., Toto, M.C.S., et Choi, J.W. (2011). A Bayesian benchmarking of the Scott-Smith model for small areas. *Journal of Statistical Computation and Simulation* (en cours d'impression, préimprimez).
- Rao, J.N.K. (2003). *Small Area Estimation*. New York : John Wiley & Sons, Inc.
- Ritter, C., et Tanner, M.A. (1992). The gibbs sampler and the gridly gibbs sampler. *Journal of the American Statistical Association*, 87, 861-868.
- Robert, C.P., et Casella, G. (1999). *Monte Carlo Statistical Methods*. New York : Springer-Verlag.
- Silvapulle, M.J., et Sen, P.K. (2006). *Constrained Statistical Inference: Inequality, Order and Shape Restrictions*. New York : John Wiley & Sons, Inc.
- Silveman, B.W. (1986). *Density Estimation*. Londres : Chapman and Hall.

Il est simple de faire une inférence prédictive bayésienne

au sujet de la moyenne de population finie de chaque petit domaine. Soit $P_i = T_i/N_i$ la proportion de la population finie pour le i^{e} domaine, où $T_i = \sum_{j=1}^{N_i} Y_{ij}$, Y_{ij} sont les réponses binaires, et N_i le nombre d'individus dans le i^{e} domaine, est supposé connu. Maintenant $T_i = t_{(s)}^i + t_{(ns)}^i$, où $t_{(s)}^i$ et $t_{(ns)}^i$ sont respectivement le total d'échantillon et le total de non-échantillon. Maintenant, sous n'importe lequel des modèles, $t_{(ns)}^i | \pi_i \sim \text{Béta-binomiale}(n_i, \pi_i)$ et $p(t_{(ns)}^i | y_s) = \int p(t_{(ns)}^i | \pi_i) p(\pi_i | y_s) d\pi_i$ où $y_s = (Y_1, \dots, Y_\ell)$. Donc, il est facile d'obtenir la densité a posteriori empirique de P_i en utilisant la méthode fondée sur l'échantillonnage.

Nandram et Sedransk (1993) ont obtenu certaines caractéristiques analytiques de P_i quand τ est connu, mais non avec la contrainte; voir aussi Nandram (1998).

Nous mentionnons une généralisation de notre modèle bayésien hiérarchique bêta-binomial restreint au modèle multinomial de Dirichlet (par exemple, Nandram 1998). Soit y_i le c -vecteur de fréquence de cellule (c'est-à-dire le nombre de personnes possédant l'un des c traits), et soit n_i la taille de l'échantillon dans le i^{e} domaine, $i = 1, \dots, \ell$.

Nous supposons que

$$y_i | \pi_i \sim \text{Multinomiale}(n_i, \pi_i), \\ \pi_i | \mu, \tau, \theta \sim \text{Dirichlet}(\mu\tau)$$

avec $\sum_{i=1}^{\ell} w_i \pi_i = \theta$. Enfin $\theta \sim \text{Dirichlet}(\mu_0 \tau_0)$, où μ_0 et τ_0 doivent être spécifiés, et indépendamment $p(\mu, \tau) = (k-1)! / (1+\tau)^2$, $0 < \mu_k < 1$, $k = 1, \dots, c$, $\sum_{k=1}^c \mu_k = 1$. Avec k contraintes, ce problème est beaucoup plus complexe, mais nous prévoyons nous y atteler. D'autres extensions à la non-réponse non ignorable (Nandram et Choi 2002) et aux tables de contingence à deux variables sont possibles.

Remerciements

Les auteurs remercient le rédacteur associé et les deux examinateurs qui ont contribué considérablement à améliorer la qualité de la présentation.

Annexe A

Preuves des lemmes 1, 2 et des théorèmes 1, 2

Preuve du lemme 1

Il s'agit d'un cas particulier d'un résultat général. En utilisant la règle de multiplication et parce que le prior est approprié, il est clair que la densité conjointe de π, μ, τ, s « s'intègre » pour donner la valeur un. Par conséquent, la densité a posteriori conjointe de π, μ, τ sachant s est appropriée.

Preuve du théorème 1

Soit $T = \{(\pi, \mu, \tau, \theta) : 0 < \pi_i < 1, i = 1, \dots, \ell, 0 < \mu < 1, \tau > 0, \theta - \omega_\ell \leq \sum_{i=1}^{\ell-1} \omega_i \pi_i \leq \theta, 0 < \theta < 1, \pi_\ell = (\theta - \sum_{i=1}^{\ell-1} \omega_i \pi_i) / \omega_\ell\}$ et $T^* = \{(\pi, \mu, \tau) : 0 < \pi_i < 1, i = 1, \dots, \ell, 0 < \mu < 1, \tau > 0\}$; notons que $T \subset T^*$. Soit $\tilde{g}(\pi, \mu, \tau | s)$ le deuxième membre de la densité a posteriori non contrainte dans (7) et $\tilde{p}(\pi^{(\ell)}, \mu, \tau, \theta | s, \phi = 0)$ le deuxième membre de la densité a posteriori dans (9). En notant que $\pi_\ell = (\theta - \sum_{i=1}^{\ell-1} \omega_i \pi_i) / \omega_\ell$, nous observons que

$$\tilde{p}(\pi^{(\ell)}, \mu, \tau, \theta | s, \phi = 0) =$$

$$\tilde{g}(\pi, \mu, \tau | s) \times \theta^{\mu_0 \tau_0 - 1} (1 - \theta)^{(1 - \mu_0) \tau_0 - 1}, (\pi, \mu, \tau, \theta) \in T.$$

Comme $\theta^{\mu_0 \tau_0 - 1} (1 - \theta)^{(1 - \mu_0) \tau_0 - 1}$ est proportionnel à la fonction de densité de la variable aléatoire bêta, nous avons

$$\int \tilde{p}(\pi^{(\ell)}, \mu, \tau, \theta | s, \phi = 0) d\pi d\mu d\tau d\theta =$$

$$A \int \tilde{g}(\pi, \mu, \tau | s) d\pi d\mu d\tau \leq A \int \tilde{g}(\pi, \mu, \tau | s) d\pi d\mu d\tau,$$

où $A = B\{\mu_0 \tau_0, (1 - \mu_0) \tau_0\}$ est la fonction bêta. En vertu du lemme 1, $\int \tilde{g}(\pi, \mu, \tau | s) d\pi d\mu d\tau < \infty$. Donc,

$$p(\pi^{(\ell)}, \mu, \tau, \theta | s, \phi = 0) \text{ est appropriée.}$$

Preuve du lemme 2 (a)

Celui-ci peut être prouvé de deux façons. La dérivée seconde de $\log\{f_2(x)\}$ est négative dans (c, d) , et donc la dérivée première, quand elle est fixée à zéro, fournit un mode unique qui est $\delta d + (1 - \delta) c$. Alternativement, comme $(X - c)/(d - c) \sim \text{Béta}(a, b)$ avec $a, b > 1$, il existe un mode unique pour $(X - c)/(d - c)$, et cela se traduit par $\delta d + (1 - \delta) c$; notons que $\delta d + (1 - \delta) c$ est un point dans (c, d) . Donc, en introduisant $\delta d + (1 - \delta) c$ par substitution dans $f_2(x)$, nous avons

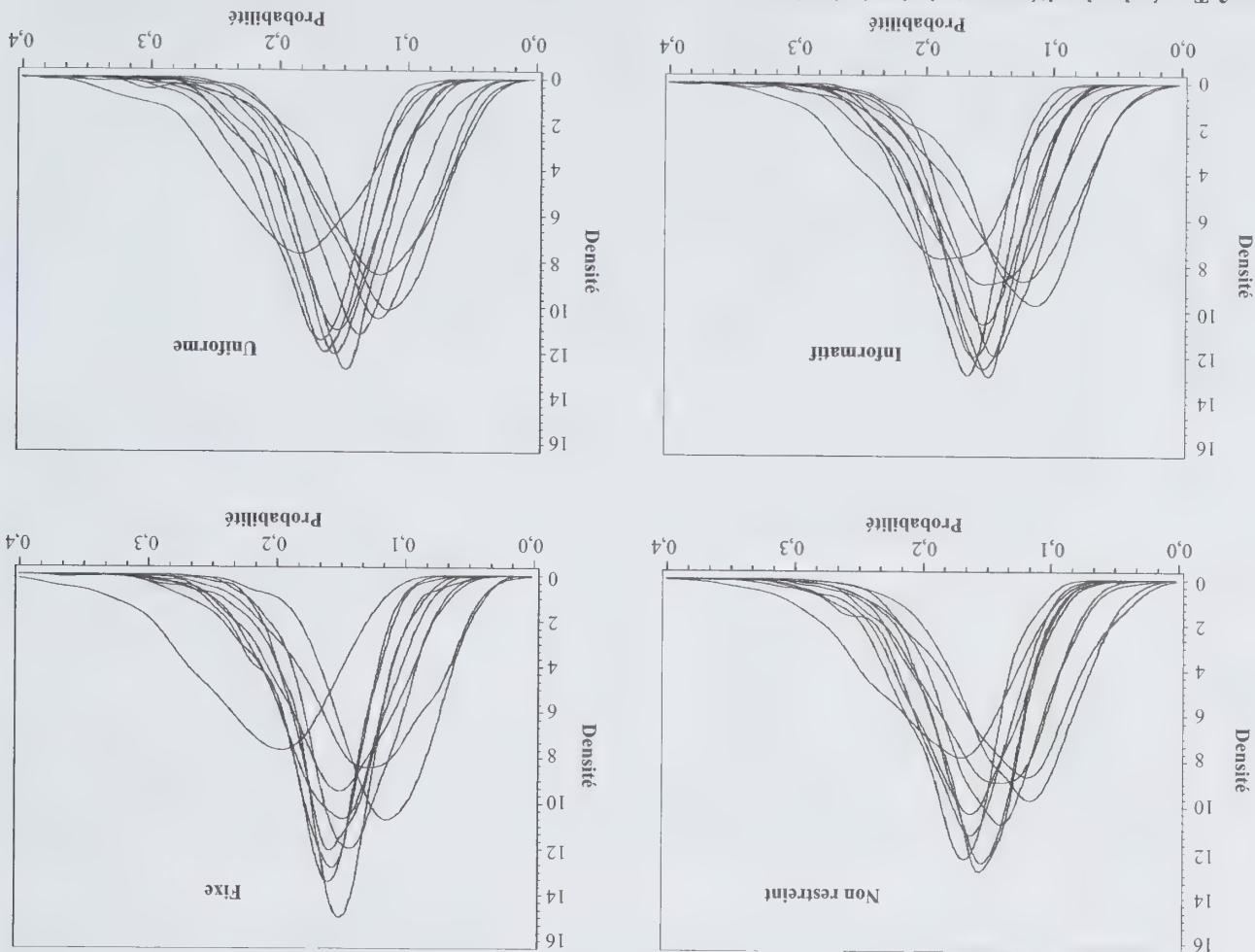
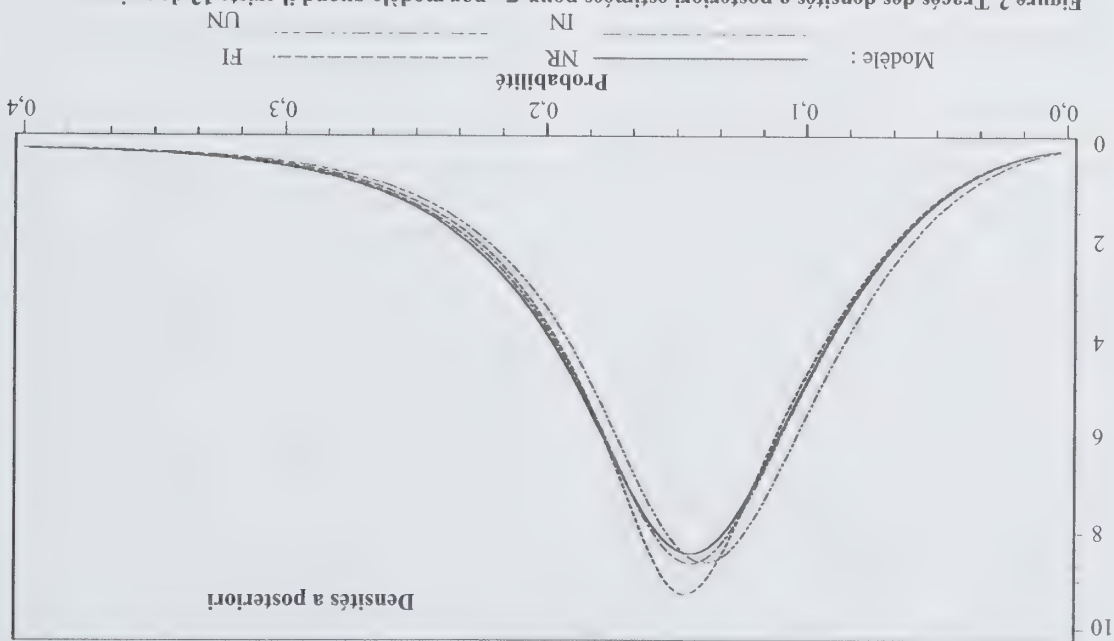
$$\sup_{c < x < d} f_2(x) = \delta^{a-1} (1 - \delta)^{b-1} / (d - c) B(a, b).$$

Preuve du lemme 2 (b)

Comme $a, b > 1$, $x \geq x - c$ et $1 - x \geq d - x$, il est vrai que

$$A^{-1} \geq D^{-1} \int_c^d (x - c)^{a+g-2} (d - x)^{b+h-2} dx,$$

où $D = (d - c)^{a+b-1} B(a, b) B(g, h) \{F_{g,h}(d) - F_{g,h}(c)\}$ et $F_{g,h}(x)$ est la fonction de répartition d'une variable aléatoire bêta standard dans $(0, 1)$. Notons que, comme $c < d$ (strictement) et $F_{g,h}(x)$ est croissante de manière monotone dans $(0, 1)$, $F_{g,h}(d) - F_{g,h}(c) > 0$ (strictement). Par comparaison avec la densité bêta généralisée [c'est-à-dire $\text{Beta}(a + g - 1, b + h - 1, c, d)$], l'intégrale est $(d - c)^{a+b+g+h-3} B(a + g - 1, b + h - 1)$. Donc,



dx densités à postériori estimées. De nouveau, nous constatons que FI est la plus élevée ; UR, FI et UN montrent une variation comparable, avec IN légèrement plus grande ; il est important de prendre la moyenne pour les comparaisons comme à la figure 2.

4. Conclusion

Nous avons étendu le modèle bêta-binomial de l'estimation sur petits domaines pour accommoder une spécification a priori d'une moyenne pondérée des probabilités de domaine. Nous avons utilisé l'approche bayésienne, qui est particulièrement séduisante pour les problèmes avec fonction de probabilité difficile à utiliser, comme dans notre application avec la contrainte de la moyenne pondérée du modèle bêta-binomial. Nous avons considéré la contrainte comme une connaissance a priori qui peut être précisée ou moins informative. L'échantillonneur de Gibbs « griddy » est utilisé pour ajuster les modèles, ce qui évite de recourir à l'échantillonneur de Metropolis-Hastings plus complexe. Nous avons élaboré une théorie qui permet un échantillonnage à partir d'une fonction de densité proportionnelle au produit d'une densité bêta-binomial tronquée et d'une densité bêta généralisée. Nous avons constaté que, dans

et inefficace.

l'ensemble, notre algorithme complet formant l'échantillonneur de Gibbs « griddy » s'exécute efficacement et rapidement. Nous avons montré qu'il pourrait y avoir des gains de précision quand de l'information supplémentaire est intégrée dans le modèle bêta-binomial. Nous avons considéré trois scénarios dans lesquels un praticien des sondages (a) ne peut spécifier aucune contrainte (modèle bêta-binomial standard pour petits domaines), (b) peut spécifier une contrainte et le paramètre complètement et (c) peut spécifier une contrainte et l'information qui peut être utilisée pour construire une distribution a priori pour le paramètre. Notre exemple d'obésité des enfants dans la National Health and Nutrition Examination Survey et l'étude par simulation ont montré que le gain de précision au-delà de (a) est dans un ordre tel que (b) est plus grand que (c). Comme les arguments algébriques exacts sont difficiles, nous avons obtenu une approximation analytique qui montre qu'il pourrait en effet y avoir un gain de précision de (b) par rapport à (a). Aux fins de comparaison, nous avons considéré un quadratisme scénario dans lequel θ possède de l'information vague et, comme prévu, il s'est avéré plutôt non intéressant

Tableau 4
Simulation : Comparaison des modèles informatif (IN) et uniforme (UN) en utilisant la moyenne à postériori (MP), la couverture (C), le biais et la moyenne de la valeur absolue du biais (B et AB), l'écart-type à postériori (ETP), la racine carrée de l'erreur quadratique moyenne (REQM) et la largeur (L) de l'intervalle de crédibilité à 95 % de π_i

ℓ	Modèle	MP	C	B	AB	ETP	REQM	L
12	IN	0,149 _{0,0012}	0,853 _{0,0112}	0,000 _{0,0003}	0,00152 _{0,00081}	0,008 _{0,0000}	0,012 _{0,0002}	0,030 _{0,0001}
	UN	0,138 _{0,0005}	0,881 _{0,0102}	-0,012 _{0,0004}	0,00038 _{0,00003}	0,011 _{0,0001}	0,016 _{0,0002}	0,042 _{0,0002}
	IN	0,153 _{0,0015}	0,833 _{0,0118}	0,003 _{0,0015}	0,00212 _{0,00103}	0,007 _{0,0006}	0,012 _{0,0015}	0,024 _{0,0015}
24	IN	0,145 _{0,0029}	0,842 _{0,0115}	-0,005 _{0,0003}	0,00012 _{0,00006}	0,008 _{0,0001}	0,010 _{0,0002}	0,030 _{0,0002}
	UN	0,150 _{0,0002}	0,828 _{0,0119}	0,000 _{0,0002}	0,00004 _{0,00000}	0,004 _{0,0000}	0,007 _{0,0001}	0,017 _{0,0001}
	IN	0,145 _{0,0003}	0,794 _{0,0128}	-0,005 _{0,0002}	0,00009 _{0,00000}	0,006 _{0,0000}	0,010 _{0,0001}	0,024 _{0,0001}
36	IN	0,150 _{0,0002}	0,828 _{0,0119}	0,000 _{0,0002}	0,00004 _{0,00000}	0,004 _{0,0000}	0,007 _{0,0001}	0,017 _{0,0001}
	UN	0,145 _{0,0003}	0,794 _{0,0128}	-0,005 _{0,0002}	0,00009 _{0,00000}	0,006 _{0,0000}	0,010 _{0,0001}	0,024 _{0,0001}
	IN	0,153 _{0,0015}	0,833 _{0,0118}	0,003 _{0,0015}	0,00212 _{0,00103}	0,007 _{0,0006}	0,012 _{0,0015}	0,024 _{0,0015}

Nota : Les deux modèles considérés sont : modèle 3 – prior informatif pour θ (IN) et modèle 4 – prior uniforme pour θ (UN). $REQM = (\pi - MP)^2 + ETP^2$. La notation a_b signifie que a est une estimation et que b est l'erreur-type.

Tableau 5
Simulation : Comparaison des quatre modèles en utilisant l'écart-type à postériori et la racine carrée de l'erreur quadratique moyenne (REQM) des π_i par domaine (D)

D	Non restreint	Fixe	Informatif	Uniforme
1	ETP 0,046 _{0,0003}	REQM 0,054 _{0,0004}	ETP 0,045 _{0,0002}	REQM 0,050 _{0,0005}
	REQM 0,055 _{0,0004}	ETP 0,044 _{0,0003}	REQM 0,044 _{0,0002}	ETP 0,047 _{0,0004}
2	ETP 0,046 _{0,0003}	REQM 0,053 _{0,0004}	ETP 0,044 _{0,0002}	REQM 0,045 _{0,0005}
	REQM 0,053 _{0,0004}	ETP 0,042 _{0,0002}	REQM 0,042 _{0,0002}	ETP 0,041 _{0,0003}
3	ETP 0,044 _{0,0002}	REQM 0,050 _{0,0004}	ETP 0,040 _{0,0002}	REQM 0,048 _{0,0003}
	REQM 0,049 _{0,0004}	ETP 0,038 _{0,0002}	REQM 0,046 _{0,0004}	ETP 0,039 _{0,0003}
4	ETP 0,040 _{0,0002}	REQM 0,046 _{0,0004}	ETP 0,039 _{0,0002}	REQM 0,048 _{0,0005}
	REQM 0,041 _{0,0002}	ETP 0,048 _{0,0004}	REQM 0,046 _{0,0004}	ETP 0,047 _{0,0005}
5	ETP 0,037 _{0,0002}	REQM 0,045 _{0,0003}	ETP 0,036 _{0,0002}	REQM 0,046 _{0,0004}
	REQM 0,038 _{0,0002}	ETP 0,040 _{0,0004}	REQM 0,037 _{0,0002}	ETP 0,048 _{0,0004}
6	ETP 0,036 _{0,0002}	REQM 0,042 _{0,0003}	ETP 0,034 _{0,0002}	REQM 0,046 _{0,0004}
	REQM 0,037 _{0,0002}	ETP 0,044 _{0,0004}	REQM 0,036 _{0,0002}	ETP 0,048 _{0,0004}
7	ETP 0,035 _{0,0002}	REQM 0,040 _{0,0004}	ETP 0,033 _{0,0002}	REQM 0,044 _{0,0004}
	REQM 0,036 _{0,0002}	ETP 0,042 _{0,0003}	REQM 0,034 _{0,0003}	ETP 0,046 _{0,0004}
8	ETP 0,034 _{0,0003}	REQM 0,039 _{0,0003}	ETP 0,032 _{0,0002}	REQM 0,042 _{0,0004}
	REQM 0,035 _{0,0003}	ETP 0,038 _{0,0003}	REQM 0,033 _{0,0003}	ETP 0,044 _{0,0004}
9	ETP 0,033 _{0,0002}	REQM 0,037 _{0,0002}	ETP 0,031 _{0,0002}	REQM 0,040 _{0,0004}
	REQM 0,034 _{0,0003}	ETP 0,036 _{0,0003}	REQM 0,030 _{0,0002}	ETP 0,042 _{0,0004}
10	ETP 0,032 _{0,0002}	REQM 0,035 _{0,0002}	ETP 0,030 _{0,0001}	REQM 0,039 _{0,0003}
	REQM 0,033 _{0,0002}	ETP 0,034 _{0,0003}	REQM 0,031 _{0,0002}	ETP 0,041 _{0,0004}
11	ETP 0,031 _{0,0002}	REQM 0,033 _{0,0002}	ETP 0,029 _{0,0001}	REQM 0,037 _{0,0003}
	REQM 0,032 _{0,0002}	ETP 0,032 _{0,0003}	REQM 0,030 _{0,0001}	ETP 0,040 _{0,0004}
12	ETP 0,030 _{0,0002}	REQM 0,031 _{0,0002}	ETP 0,028 _{0,0001}	REQM 0,036 _{0,0003}
	REQM 0,031 _{0,0002}	ETP 0,030 _{0,0003}	REQM 0,029 _{0,0001}	ETP 0,042 _{0,0004}

Note : Quatre modèles sont : modèle 1 – non restreint (NR) : modèle 2 – θ fixe (FI) : modèle 3 – prior informatif pour θ (IN) : modèle 4 – prior uniforme pour θ (UN). $REQM = (\pi - MP)^2 + ETP^2$. La notation a_b signifie que a est une estimation et que b est l'erreur-type, ici. 12 domaines sont utilisés et les tailles d'échantillon originales sont divisées par 2.

Tableau 2

Simulation : Comparaison des quatre modèles en utilisant la couverture (C), le biais et la moyenne de la valeur absolue du biais (B) et AB), l'écart-type a posteriori (ETP), la racine carrée de l'erreur quadratique moyenne (REQM) et la largeur (L) de l'intervalle de crédibilité à 95 % de π_i

ℓ	Modèle	C	B	AB	ETP	REQM	L
12	NR	0,960 ^{0,0018}	-0,002 ^{0,0003}	0,0231 ^{0,00016}	0,033 ^{0,0001}	0,043 ^{0,0001}	0,125 ^{0,0003}
	FI	0,961 ^{0,0018}	-0,000 ^{0,0003}	0,0219 ^{0,00020}	0,031 ^{0,0001}	0,040 ^{0,0001}	0,118 ^{0,0003}
	IN	0,946 ^{0,0021}	0,005 ^{0,0003}	0,0275 ^{0,00066}	0,032 ^{0,0001}	0,043 ^{0,0001}	0,122 ^{0,0002}
	NR	0,956 ^{0,0019}	-0,000 ^{0,0003}	0,0261 ^{0,00019}	0,032 ^{0,0001}	0,042 ^{0,0001}	0,122 ^{0,0003}
24	NR	0,957 ^{0,0003}	-0,001 ^{0,0002}	0,0229 ^{0,00012}	0,031 ^{0,0000}	0,041 ^{0,0001}	0,119 ^{0,0002}
	FI	0,957 ^{0,0013}	-0,000 ^{0,0002}	0,0224 ^{0,00013}	0,030 ^{0,0000}	0,040 ^{0,0001}	0,116 ^{0,0002}
	IN	0,943 ^{0,0015}	0,006 ^{0,0002}	0,0252 ^{0,00058}	0,030 ^{0,0000}	0,041 ^{0,0001}	0,116 ^{0,0001}
	NR	0,952 ^{0,0014}	-0,000 ^{0,0002}	0,0236 ^{0,00012}	0,031 ^{0,0002}	0,041 ^{0,0002}	0,118 ^{0,0005}
36	NR	0,960 ^{0,0010}	-0,001 ^{0,0001}	0,0224 ^{0,00009}	0,030 ^{0,0000}	0,040 ^{0,0001}	0,117 ^{0,0001}
	FI	0,961 ^{0,0010}	-0,000 ^{0,0001}	0,0218 ^{0,00009}	0,030 ^{0,0000}	0,039 ^{0,0001}	0,115 ^{0,0001}
	IN	0,948 ^{0,0012}	0,005 ^{0,0002}	0,0224 ^{0,00009}	0,030 ^{0,0000}	0,040 ^{0,0001}	0,114 ^{0,0001}
	NR	0,957 ^{0,0001}	-0,000 ^{0,0001}	0,0228 ^{0,00010}	0,030 ^{0,0000}	0,040 ^{0,0001}	0,116 ^{0,0001}

Nota : Quatre modèles sont : modèle 1 non restreint (NR) ; modèle 2 - θ fixe (FI) ; modèle 3 - prior informatif pour θ (IN) ; modèle 4 - prior uniforme pour θ (UN). REQM = $(\pi - MP)^2 + ETP^2$. La notation a_b signifie que a est une estimation et que b est l'erreur-type.

Dans la plupart des applications, la valeur exacte de θ est inconnue. Par conséquent, les ETP des π_i dans la situation où θ est supposé connu, sous-estime vraisemblablement les ETP réels. Donc, nous étudions les écarts des ETP de IN et UN par rapport à ceux de FI, et nous calculons les ratios, $R_1 = ETP^{IN} / ETP^{FI}$ et $R_2 = ETP^{UN} / ETP^{FI}$. Dans le tableau 3, nous présentons les sommes à cinq chiffres de ces ratios selon la taille d'échantillon. La plupart des ratios sont situés autour de 1 (c'est-à-dire intervalle interquartile) avec une certaine tendance à être plus grands que 1. (Notons que les maxima à $\ell = 12$ et $\ell = 24$ sont des valeurs aberrantes, peut-être dues à de mauvais échantillons simulés.) Donc, dans l'ensemble, les ETP sous IN et UN ne sont pas beaucoup plus grands que sous FI.

Tableau 3
Simulation : Une étude de l'écart-type a posteriori (ETP) de π_i en utilisant les grandes sommes à cinq chiffres des ratios, R_1 et R_2 , selon la taille d'échantillon

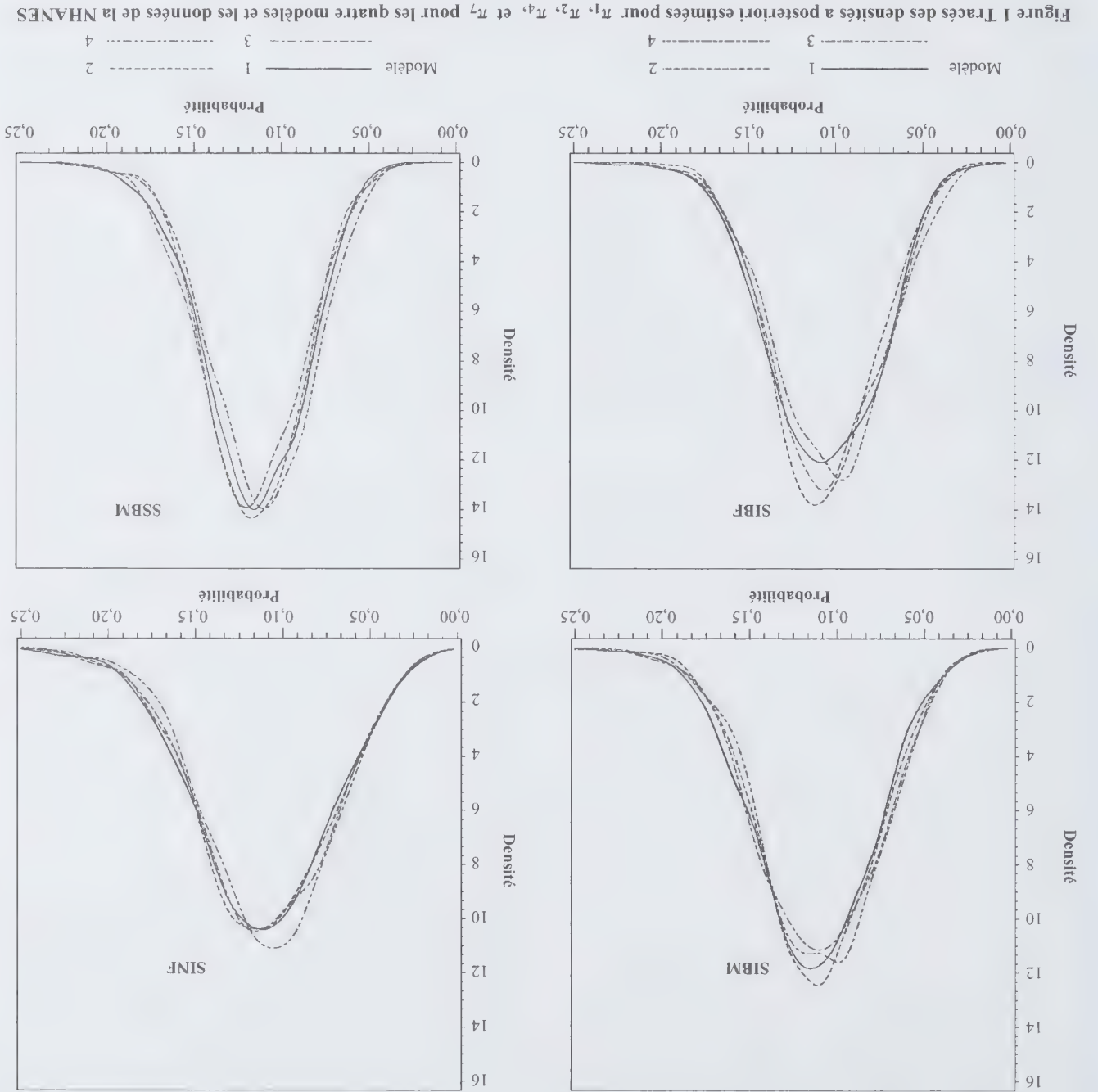
ℓ	Ratio	Min	Q_1	Méd	Q_3	Max
12	R_1	0,673	0,972	1,032	1,091	5,329
	R_2	0,022	0,984	1,034	1,086	85,370
24	R_1	0,019	0,965	1,005	1,047	16,017
	R_2	0,024	0,979	1,014	1,049	486,960
36	R_1	0,690	0,962	0,998	1,034	1,236
	R_2	0,837	0,979	1,011	1,044	1,243

Nota : $R_1 = ETP^{IN} / ETP^{FI}$ et $R_2 = ETP^{UN} / ETP^{FI}$. Les cinq grandes sommes sont le minimum (min), le premier quartile (Q_1), la médiane (méd), le troisième quartile (Q_3) et le maximum (max).

Dans le tableau 4, nous étudions l'estimation de θ pour deux modèles pertinents IN et UN. Pour les deux modèles, les probabilités de couverture sont plus faibles que la valeur nominale, et la couverture pour UN est plus faible que l'intervalle pour IN. Le biais est faible pour les deux modèles, positif pour IN et négatif pour UN. Sauf pour

Nous pouvons voir une variation importante entre les densités pour les 1 000 simulations exécutées par modèle. figure 3, nous présentons un échantillon systématique de dix UN est déplacé légèrement vers la gauche de IN. Dans la laires. FI est la densité la plus élevée et UN la plus courte. les données sur l'IMC. De nouveau, les queues sont similaires pour $\ell = 12$. Nous obtenons les mêmes résultats que pour (Parzen-Rosenblatt) moyennées sur les 1 000 exécutions la figure 2, nous présentons les densités a posteriori estimées utilisons l'estimateur de la densité de Parzen-Rosenblatt. À et nous comparons les quatre modèles. De nouveau, nous Nous étudions la densité a posteriori de π_1 pour $\ell = 12$, diminuent à mesure que le nombre de domaines augmente.

Dans le tableau 5, nous présentons des résultats plus légèrement meilleurs résultats. détails (c'est-à-dire par domaine) pour le cas où le nombre de domaines est 12. Pour montrer des gains supplémentaires de précision, nous avons réduit la taille d'échantillon de moitié [c'est-à-dire nous avons tiré les tailles d'échantillon uniformément dans l'intervalle (12 ; 75)]. Nous présentons l'écart-type a posteriori et la racine carrée de l'erreur quadratique moyenne a posteriori, moyennées sur les exécutions de simulation. De nouveau, les erreurs-types sont présentes. Nous notons que tous les contenus de probabilité (non présentes) sont au moins égaux à la valeur nominale de 95 %. Les erreurs-types numériques sont faibles dans tous les cas. Les ETP et les REQM sont dans le bon ordre. Notons que, parce que les tailles d'échantillon sont classées par ordre de la plus petite à la plus grande, les ETP et les REQM diminuent à mesure que le nombre de domaines augmente.



Dans le tableau 2, nous étudions les estimations des probabilités de petits domaines. Il est commode d'utiliser des noms abrégés des quatre modèles pour nos discussions. Pour IN, les MP sont proches de la valeur nominale de 0,15, mais pour UN, les MP sont plus petites que la valeur nominale, particulièrement pour NR à $\ell = 12$. Nous observons que la couverture pour tous les modèles NR, FI et UN est toujours plus grande que la valeur nominale de 95 %, mais pour le modèle IN, la couverture est plus faible que la valeur nominale de 95 %. Une différence similaire existe pour le biais ; bien que le biais soit faible pour tous les modèles, les

modèles NR, FI (la valeur spécifiée de θ est 0,15) et UN ont un biais négatif, mais IN a un biais positif. Sauf pour la plupart principalement semblables et les REQM ont les mêmes caractéristiques ; il existe quelques différences à $\ell = 12$. Les quatre modèles deviennent similaires à mesure que ℓ augmente ; quand ℓ est grand, il semble que notre méthode ne soit plus nécessaire. Cependant, de nouveau, le gain de précision semble être par ordre croissant FI, IN, UN et NR.

Nous étudions également très brièvement le paramètre de nuisance θ . Nous notons que la moyenne pondérée des estimateurs directs des petits domaines est 0,136 (plus exactement 0,1355599). Quand θ est maintenu fixe à 0,136, Quand θ possède le prior informatif, la moyenne pondérée des moyennes a posteriori est 0,136, et pour θ , la MP est 0,136, l'ETP est 0,008, et un intervalle DPM à 95 % pour θ est (0,122 ; 0,152). Quand θ a le prior uniforme, la moyenne pondérée des moyennes a posteriori est 0,132, et pour θ la MP est 0,131, l'ETP est 0,011, et un intervalle DPM à 95 % pour θ est (0,110 ; 0,151). Cela montre les déficiences de la loi a priori uniforme que nous utilisons uniquement aux fins de comparaison. Il convient de souligner que $\mu_1, \dots, \mu_{\ell-1}$ et θ sont calculés pour commencer. Puis μ_ℓ est obtenu par soustraction. Cela est fait à chaque itération de l'échantillonneur de Gibbs. Ensuite, les valeurs sommatrices a posteriori pour $\sum_{i=1}^{\ell} \omega_i \pi_i$ et θ sont calculées. Donc, il y aura de légères discordances qui sont dues à l'arrondissement.

Enfin, nous avons sélectionné les quatre plus petits domaines pour comparer les densités a posteriori des probabilités. Nous avons utilisé l'estimateur à noyau de la densité de probabilité de Parzen-Rosenblatt pour estimer les densités a posteriori ; voir Silverman (1986) pour des détails. La figure 1 compare les densités a posteriori estimées pour les quatre modèles. Il est intéressant de noter que, à mesure que les tailles des domaines augmentent, les quatre modèles se rapprochent. En outre, dans tous les cas, les queues de distribution dans chaque panneau sont très semblables ; les différences entre ces distributions tiennent aux intervalles modaux (c'est-à-dire intervalle contenant le mode), et leurs hauteurs. Comme prévu, la densité a posteriori correspondant au modèle non restreint est la plus courte, simplement parce qu'elle a plus de variabilité. Le modèle 4 voit sa densité a posteriori déplacée vers la gauche et est légèrement bimodal pour le domaine le plus petit. Donc, l'inférence au sujet des modes de ces distributions sera différente. Mais l'inférence faisant intervenir les queues ne sera pas si différente ; sauf pour le modèle 4, les intervalles de crédibilité à 95 % seront similaires.

3.2 Étude en simulation

Nous utilisons une étude en simulation pour évaluer les propriétés statistiques de notre méthode. Nous voulons voir si le gain de précision persiste et comment les estimateurs des probabilités sont déplacés. Nous étudions aussi les propriétés fréquentistes des estimateurs des probabilités. Dans la description de la simulation, il est commode d'utiliser les noms abrégés des modèles qui sont NFR (modèle 1, non restreint), FI (modèle 2, θ fixe), IN (modèle 3,

$$\begin{aligned} \pi_i &\sim \text{Bêta}\{\mu_0 \tau_0, (1 - \mu_0) \tau_0\}, i = 1, \dots, \ell. \\ \ell \pi_i &\text{ jusqu'à } \theta_0 - w_\ell \leq \sum_{i=1}^{\ell} \omega_i \pi_i \leq \theta_0 ; \text{ ensemble } \pi_\ell = (\theta_0 - \sum_{i=1}^{\ell} \omega_i \pi_i) / w_\ell. \text{ Puis, nous avons généré} \\ s_i &\sim \text{Binomiale}(n_i, \pi_i). \end{aligned}$$

Pour exécuter cette tâche, nous avons tiré des ensembles de $\ell \pi_i$ jusqu'à $\theta_0 - w_\ell \leq \sum_{i=1}^{\ell} \omega_i \pi_i \leq \theta_0$; ensemble $\pi_\ell = (\theta_0 - \sum_{i=1}^{\ell} \omega_i \pi_i) / w_\ell$. Puis, nous avons généré $s_i \sim \text{Binomiale}(n_i, \pi_i)$. Nous avons généré un nombre de domaines dans les données de la NHANES. Nous avons tiré les tailles d'échantillon à partir d'une densité de probabilité uniforme dans (25 ; 150), de nouveau pour refléter les données de la NHANES. Premièrement, nous avons généré

Nous avons généré 1 000 ensembles de données de cette manière pour chacun des $\ell = 12, 24, 36$. Puis, nous avons ajusté les quatre modèles (un modèle non restreint et trois modèles restreints). Le processus est très rapide (c'est-à-dire, pour les tailles d'échantillon de 12, 24, 30, il y a eu respectivement 22, 90 et 153 rejets dans les 1 000 échantillons). Nous avons ajusté chaque ensemble de données en utilisant des échantillons aléatoires pour le modèle non restreint, et l'échantillonneur de Gibbs « gridly » pour les modèles restreints. Nous avons ajusté les 1 000 ensembles de données en une ou deux heures sur notre ordinateur Alpha 2 \times 833 MHz.

Pour ces 1 000 simulations, nous étudions la MP, la couverture (C), le biais (B), l'ETP, la REQM et la largeur (L) des intervalles de crédibilité à 95 %. Pour chaque domaine, nous calculons le biais $MP - \pi_i$, puis nous calculons la moyenne de ces valeurs sur tous les domaines et toutes les exécutions des simulations, et nous appelons maintenant cette quantité B. Associé à B, nous avons aussi calculé AB, la moyenne de $|MP - \pi_i|$. De même, nous avons calculé

$$REQM = \sqrt{(MP - \pi_i)^2 + ETP^2}$$

pour chaque domaine et chaque exécution de la simulation et nous avons calculé la moyenne de ces valeurs sur tous les domaines et toutes les exécutions de simulation. Notons que les probabilités réelles, π_i , sont connues par conception. Nous obtenons la couverture (C) en calculant la proportion de tous les intervalles contenant la valeur réelle de π_i sur tous les domaines et toutes les exécutions de simulation. Nous obtenons aussi la moyenne des largeurs des intervalles de crédibilité à 95 %. Les erreurs-types numériques sont obtenues pour toutes les quantités.

modèle bêta-binomial sont biaisés si le modèle spécifié est

incorrect.

Nous avons pris $\mu_0 = 0,136$, la proportion globale

d'échantillon, et $\tau_0 = 959$, la taille totale d'échantillon.

Des choix moins optimistes peuvent être utilisés. Par exem-

ple, $\tau_0 = 100$, disons ; mais ce choix fait peu de différence.

Cependant, il convient de souligner qu'utiliser les données

observées pour spécifier la distribution a priori peut dimi-

nuer artificiellement la variance a posteriori. Habituel-

lement, un praticien des sondages a une spécification ap-

propriée venant d'une enquête antérieure ou d'un recense-

ment. On ne peut pas spécifier des valeurs pour μ_0 et τ_0

qui sont entièrement déraisonnables et qui créeront des biais

énormes. Ici, τ_0 est une taille d'échantillon a priori et μ_0

est une moyenne a priori de θ . Cette méthode permet une

valeur raisonnable pour θ ; nous ajoutons essentiellement

un degré d'incertitude au sujet de la connaissance de la

combinaison linéaire. Donc, ces spécifications ne sont pas

déraisonnables.

Nous avons appliqué notre méthode telle que décrite pour les quatre scénarios. Dans les autres colonnes du ta-bleau 1, nous étudions les estimations des probabilités de petits domaines. Nous présentons la moyenne a posteriori (MP), l'écart-type a posteriori (ETP), la racine carrée de l'erreur quadratique moyenne (REQM), et les intervalles de crédibilité DPM à 95 % (HPD) (Int) de π_i par domaine (D) pour les données de la NHANES

Tableau 1

Comparaison des quatre modèles en utilisant la moyenne a posteriori (MP), l'écart-type a posteriori (ETP), la racine carrée de l'erreur quadratique moyenne (REQM), et les intervalles de crédibilité DPM à 95 % (HPD) (Int) de π_i par domaine (D) pour les données de la NHANES

D	s	n	$\hat{\pi}$	MP	ETP	REQM	Int	MP	ETP	REQM	Int
---	---	---	-------------	----	-----	------	-----	----	-----	------	-----

Modèle 1											
1	4	47	0,085	0,114	0,033	0,044	(0,051 ; 0,179)	0,111	0,032	0,041	(0,049 ; 0,170)
2	2	29	0,069	0,112	0,037	0,057	(0,042 ; 0,183)	0,111	0,036	0,055	(0,041 ; 0,178)
3	10	44	0,227	0,175	0,044	0,068	(0,100 ; 0,264)	0,177	0,041	0,065	(0,108 ; 0,260)
4	5	62	0,081	0,107	0,030	0,040	(0,047 ; 0,159)	0,107	0,027	0,038	(0,054 ; 0,160)
5	10	74	0,135	0,134	0,030	0,030	(0,077 ; 0,194)	0,134	0,028	0,030	(0,080 ; 0,190)
6	12	69	0,174	0,158	0,036	0,039	(0,089 ; 0,227)	0,155	0,031	0,036	(0,095 ; 0,214)
7	8	79	0,101	0,116	0,028	0,031	(0,065 ; 0,173)	0,115	0,027	0,030	(0,065 ; 0,166)
8	5	62	0,081	0,107	0,030	0,040	(0,052 ; 0,169)	0,105	0,029	0,038	(0,042 ; 0,153)
9	28	123	0,228	0,196	0,036	0,048	(0,129 ; 0,262)	0,196	0,032	0,045	(0,131 ; 0,253)
10	10	111	0,090	0,106	0,026	0,030	(0,059 ; 0,155)	0,105	0,024	0,028	(0,061 ; 0,150)
11	16	122	0,131	0,132	0,026	0,026	(0,083 ; 0,183)	0,130	0,023	0,023	(0,090 ; 0,179)
12	20	137	0,146	0,144	0,026	0,026	(0,094 ; 0,194)	0,141	0,022	0,023	(0,100 ; 0,184)
Modèle 3											
1	4	47	0,085	0,042	(0,044 ; 0,169)	0,109	0,032	0,040	(0,050 ; 0,172)	0,091	(0,091 ; 0,189)
2	2	29	0,069	0,037	(0,039 ; 0,179)	0,108	0,036	0,053	(0,037 ; 0,173)	0,072	(0,091 ; 0,255)
3	10	44	0,227	0,175	(0,093 ; 0,260)	0,170	0,044	0,072	(0,048 ; 0,164)	0,030	(0,067 ; 0,184)
4	5	62	0,081	0,106	(0,050 ; 0,160)	0,103	0,030	0,038	(0,048 ; 0,164)	0,030	(0,067 ; 0,184)
5	10	74	0,135	0,134	(0,077 ; 0,189)	0,129	0,030	0,030	(0,067 ; 0,184)	0,030	(0,067 ; 0,184)
6	12	79	0,174	0,156	(0,090 ; 0,217)	0,151	0,036	0,043	(0,087 ; 0,222)	0,029	(0,051 ; 0,149)
7	8	69	0,101	0,118	(0,062 ; 0,171)	0,111	0,028	0,029	(0,061 ; 0,167)	0,025	(0,051 ; 0,149)
8	5	62	0,081	0,107	(0,051 ; 0,165)	0,102	0,030	0,036	(0,050 ; 0,159)	0,025	(0,051 ; 0,149)
9	28	123	0,228	0,195	(0,138 ; 0,265)	0,189	0,035	0,032	(0,123 ; 0,253)	0,025	(0,051 ; 0,149)
10	10	111	0,090	0,107	(0,062 ; 0,156)	0,104	0,025	0,029	(0,051 ; 0,149)	0,025	(0,051 ; 0,149)
11	16	122	0,131	0,132	(0,086 ; 0,179)	0,126	0,025	0,025	(0,083 ; 0,179)	0,025	(0,083 ; 0,179)
12	20	137	0,146	0,143	(0,095 ; 0,191)	0,137	0,025	0,027	(0,091 ; 0,189)	0,027	(0,091 ; 0,189)
Modèle 4											
1	4	47	0,085	0,040	(0,044 ; 0,169)	0,109	0,032	0,040	(0,050 ; 0,172)	0,091	(0,091 ; 0,189)
2	2	29	0,069	0,037	(0,039 ; 0,179)	0,108	0,036	0,053	(0,037 ; 0,173)	0,072	(0,091 ; 0,255)
3	10	44	0,227	0,175	(0,093 ; 0,260)	0,170	0,044	0,072	(0,048 ; 0,164)	0,030	(0,067 ; 0,184)
4	5	62	0,081	0,106	(0,050 ; 0,160)	0,103	0,030	0,038	(0,048 ; 0,164)	0,030	(0,067 ; 0,184)
5	10	74	0,135	0,134	(0,077 ; 0,189)	0,129	0,030	0,030	(0,067 ; 0,184)	0,030	(0,067 ; 0,184)
6	12	79	0,174	0,156	(0,090 ; 0,217)	0,151	0,036	0,043	(0,087 ; 0,222)	0,029	(0,051 ; 0,149)
7	8	69	0,101	0,118	(0,062 ; 0,171)	0,111	0,028	0,029	(0,061 ; 0,167)	0,025	(0,051 ; 0,149)
8	5	62	0,081	0,107	(0,051 ; 0,165)	0,102	0,030	0,036	(0,050 ; 0,159)	0,025	(0,051 ; 0,149)
9	28	123	0,228	0,195	(0,138 ; 0,265)	0,189	0,035	0,032	(0,123 ; 0,253)	0,025	(0,051 ; 0,149)
10	10	111	0,090	0,107	(0,062 ; 0,156)	0,104	0,025	0,029	(0,051 ; 0,149)	0,025	(0,051 ; 0,149)
11	16	122	0,131	0,132	(0,086 ; 0,179)	0,126	0,025	0,025	(0,083 ; 0,179)	0,025	(0,083 ; 0,179)
12	20	137	0,146	0,143	(0,095 ; 0,191)	0,137	0,025	0,027	(0,091 ; 0,189)	0,027	(0,091 ; 0,189)

Nota : Les quatre modèles sont : modèle 1 – pas de contrainte ; modèle 2 – θ fixe ; modèle 3 – prior informatif pour θ ; modèle 4 – prior uniforme

pour θ . Les domaines sont formés par croisement de l'école (secondaire inférieur – SI, secondaire supérieur – SS), la race (blanche – B, noire – N, mexicaine – M) et le sexe (masculin – M, féminin – F). Donc, les domaines sont : 1-SIBM, 2-SINF, 3-SIMM, 4-SIBF, 5-SINM, 6-SIMF, 7-SSBM, 8-SSNF, 9-SSMM, 10-SSBF, 11-SSNM, 12-SSMF (par exemple, le premier domaine constitue des garçons blancs au cycle secondaire inférieur). n est le nombre d'adolescents et s est le nombre d'adolescents obèses dans chaque domaine. Les données proviennent des 35 plus grands comités des E-U. Une estimation de la probabilité globale est $130 / 959 \approx 0,136$, et pour le premier domaine $\hat{p} = 4 / 47 = 0,085$; les erreurs-types numériques sont toutes plus petites que 0,001 ; REQM = $\sqrt{(\hat{\pi} - MP)^2 + ETP^2}$.

Algorithme

(a) Tirer $U \sim \text{Uniforme}(0, 1)$ et poser

$$\pi = F_{-1}^{-g,h}(d) \{UF_{g,h}^g(d) + (1 - U)F_{g,h}^g(c)\},$$

(b) Tirer $V \sim \text{Uniforme}(0, 1)$. Si

$$V \leq \frac{1}{1 - c} \left(\frac{d - c}{d - c} \right)^{a+b-2} \left(\frac{d - \pi}{d - c} \right)^{b-1} \left(\frac{1 - \delta}{1 - c} \right)^{a-1},$$

accepter π , sinon aller à (a).

Comme les tailles d'échantillon binomial sont arrangées par ordre croissant, dans toute application il sera vrai que $a, b > 1$ et $g, h > 0$ (éventuellement plus grand que 1). Donc, l'algorithme fonctionnera. En effet, dans tous nos exemples (un présente ici) et exercices de simulation, l'algorithme s'exécute très rapidement.

Maintenant, nous montrons comment tirer $\pi_i, i = 1, \dots, \ell$, et θ . Pour π_i ,

$$P(\pi_i | \pi_{(i,\ell)}, \theta, \mu, \tau, s, \phi = 0)$$

$\propto \pi_{a_i-1}^{b_i-1} (1 - \pi_i)^{b_i-1} (d_i - \pi_i)^{a_i-1} (d_i - c_i)^{b_i-1} (d_i - \pi_i)^{a_i-1} c_i^{b_i-1}$,
où $\pi_{(i,\ell)}$ est le vecteur contenant les éléments de π , sauf pour π_i et π_{ℓ} , et $a_i = s_i + \mu\tau, b_i = f_i + (1 - \mu)\tau, i = 1, \dots, \ell$,

$$c_i = \left(\theta - \sum_{j=1}^{i-1} \omega_j \pi_j - \omega_i \right) / \omega_i,$$

$$d_i = \left(\theta - \sum_{j=1, j \neq i}^{\ell} \omega_j \pi_j \right) / \omega_i, i = 1, \dots, \ell - 1.$$

Appliquer le théorème à $P(\pi_i | \pi_{(i)}, \theta, \mu, \tau, s), a_i > 1, b_i > 1, i = 1, \dots, \ell - 1$.

Pour θ , nous avons

$$P(\theta | \pi, \mu, \tau, s, \phi = 0)$$

$$\propto \theta^{\mu_0} \tau_0^{\tau_0-1} (1 - \theta)^{(1-\mu_0)\tau_0-1} (\theta - c)^{a-1} (d - \theta)^{b-1}, c > \theta > d,$$

où

$$c = \sum_{i=1}^{\ell} \omega_i \pi_i, d = \omega_{\ell} + \sum_{i=1}^{\ell-1} \omega_i \pi_i.$$

De nouveau, appliquer le théorème, $a_i > 1, b_i > 1$.

Quand θ est entièrement spécifié (c'est-à-dire θ n'est pas aléatoire), nous ne devons pas tirer θ . Cependant, quand $\theta \sim \text{Uniforme}(0, 1)$ a priori ($\mu_0 = 1/2, \tau_0 = 2$), nous avons une simplification. Dans ce cas,

$$\theta | \pi_{(i)}, \mu, \tau, s, \phi = 0 \sim \text{GenB\acute{e}ta}(a_i, b_i, c, d)$$

et $\theta = c + (d - c)X$, où $X \sim \text{B\acute{e}ta}(a_i, b_i)$, a la densité

requise.

Pour les modèles restreints ainsi que non restreints, nous utilisons 10 000 itérations pour faire l'inférence a posteriori au sujet des probabilités binomiales, π_i . Sous le modèle non

3.1 Exemple

Nous avons utilisé des données de la troisième National Health and Nutrition Examination Survey (NHANES) pour illustrer notre méthode. Nous avons étudié l'indice de masse corporelle des adolescents et nous avons des données sur l'échantillon obtenu. Les domaines (petits domaines) sont formés par croisement de l'ethnicité (blanche, noire, mexicaine) et du sexe (masculin, féminin). Nous avons classé les adolescents selon qu'ils étaient au secondaire inférieur ou supérieur au moment de l'enquête. Donc, il existe 12 petits domaines. Les données sont présentées dans les quatre premières colonnes du tableau 1 par domaine. Notons que les domaines SIBM, SINP, SIBF et SSNF sont relativement peu peuplés avec 4, 2, 5, 5 adolescents obèses respectivement ; pour les 12 domaines, l'échantillon consiste en 959 adolescents dont 130 sont obèses (c'est-à-dire la proportion globale des individus est de 0,136 environ). Dans la colonne 4 du tableau 1, nous avons présenté les estimations directes par domaine et ces estimations varient de 0,069 à 0,228. Les estimations pour les domaines les plus petits ne sont pas fiables. En outre, quand les modèles bêta-binomiaux sont utilisés, ces estimations régressent vers la moyenne globale d'échantillon de 0,136, ce qui crée un biais éventuel. Notre méthode devrait augmenter la précision au-delà du modèle non restreint, parce que le modèle restreint utilise plus d'information au sujet de la somme pondérée. Clairement, les prédicteurs basés sur soit le modèle restreint, soit le

3. Études numériques

À la section 3.1, nous décrivons un exemple pour illustrer les principales caractéristiques de la contrainte. À la section 3.2, nous décrivons une étude en simulation pour montrer les propriétés fréquentistes des estimateurs de Bayes, et nous donnons un aperçu plus approfondi des différences entre les quatre scénarios. Notons de nouveau que, quand nous exécutons les calculs, il est commode de classer les domaines par taille de façon que le domaine le plus grand soit le dernier.

ordinateur alpha 2 × 833 MHz.

Donc, nous obtenons les échantillons de $\pi_1, \dots, \pi_{\ell-1}$

et nous posons

$$\pi_{\ell} = \frac{\omega_{\ell}}{\left(\theta - \sum_{i=1}^{\ell-1} \omega_i \pi_i \right)}$$

pour compléter le vecteur π_1, \dots, π_{ℓ} . Autrement dit, la contrainte est obtenue exactement. La densité a posteriori conditionnelle de θ est

$$p(\theta | \pi^{(\ell)}, \mu, \tau, s, \phi = 0) \propto \left\{ \theta - \sum_{i=1}^{\ell-1} \omega_i \pi_i \right\}_{s'+\tau+1}^{\ell} \left\{ \omega_{\ell} + \sum_{i=1}^{\ell-1} \omega_i \pi_i - \theta \right\}_{f'+(1+\ell)(\mu-\tau)+1}^{f'+(1+\ell)(\mu-\tau)+1} \times \theta^{\mu_0 \tau_0 - 1} (1 - \theta)^{(\ell - \mu_0) \tau_0 - 1}, \quad (17)$$

où

$$\sum_{i=1}^{\ell-1} \omega_i \pi_i < \theta < \omega_{\ell} + \sum_{i=1}^{\ell-1} \omega_i \pi_i.$$

La densité a posteriori conditionnelle conjointe de μ et τ est

$$p(\mu, \tau | \pi^{(\ell)}, \theta, s, \phi = 0)$$

$$\propto \frac{q^{\mu \tau} p^{(1-\mu)\tau}}{B(\mu \tau, (1-\mu)\tau)} \times \frac{(1+\tau)^{\ell}}{1}, \quad (18)$$

$$0 < \mu < 1, \tau > 0, q = \prod_{i=1}^{\ell} \pi_i, r = \prod_{i=1}^{\ell} (1 - \pi_i).$$

Pour exécuter l'échantillonneur de Gibbs, nous devons

tirer des échantillons de (16), (17) et (18), tour à tour, jusqu'à la convergence. Nous tirons μ, τ de $p(\mu, \tau | \pi^{(\ell)}, \theta, s)$ de manière semblable au tirage de $p(\mu, \tau | \pi^{(\ell)})$ dans le modèle non restreint. Il est plus difficile de tirer un échantillon de (16) et (17). Cependant, nous utilisons essentiellement la même méthode pour tirer les échantillons de la densité a posteriori conditionnelle de $\pi_i, i = 1, \dots, \ell - 1$, obtenue de (16) et θ à partir de (17) qui sont toutes deux proportionnelles au produit de deux fonctions de densité, l'une étant une densité bêta tronquée et l'autre, une densité bêta généralisée. Ensuite, nous élaborons une certaine théorie pour tirer un échantillon d'une telle densité. Pour cela, nous énonçons et prouvons le lemme 2 et le théorème 2.

La fonction de densité d'intérêt est

$$f(x) = Af_1(x)f_2(x), 0 \leq c < x < d \leq 1, \quad (19)$$

où

$$f_1(x) = \frac{x^{s-1} (1-x)^{h-1}}{\int_0^1 x^{s-1} (1-x)^{h-1} dx}, c < x < d, g, h > 0, \quad (20)$$

$$f_2(x) = (x - c)^{a-1} (d - c)^{a+b-1} \{ (d - c)^{a+b-1} B(a, b) \},$$

$$c < x < d, a, b > 1, \quad (21)$$

et, naturellement,

$$A = 1 / \int_c^d f_1(x) f_2(x) dx. \quad (22)$$

Il convient de souligner que nous ne supposons pas que $g, h > 1$. Si cela était le cas, $f_1(x)$ et $f_2(x)$ seraient toutes deux logconcaves, ce qui rendrait $f(x)$ logconcave, et, dans ce cas, on peut tirer un échantillon de $f(x)$ en utilisant l'échantillonneur par rejet adaptif (ARS, Gilks et Wild 1992). Nous fournissons un algorithme spécialisé pour tirer un échantillon de $f(x)$ qui n'est pas logconcave. Même si $f_1(x)$ était logconcave (c'est-à-dire $g, h > 1$) cet algorithme spécialisé serait encore supérieur à l'échantillonneur ARS, parce que ce dernier est un algorithme d'usage général; voir Robert et Casella (1999, page 59). Notre algorithme requiert moins de calcul et ne nécessite pas la logconcavité; même en présence de logconcavité, l'échantillonneur ARS peut donner de mauvais résultats dans les queues de la fonction de densité.

Lemme 2 *Considérons les fonctions de densité $f_1(x)$ et $f_2(x)$ avec $a, b > 1$.*

(a) *Alors*

$$\sup_{c < x < d} f_2(x) = \frac{\delta^{a-1} (1 - \delta)^{b-1}}{(d - c) B(a, b)}, \delta = (a - 1) / (a + b - 2).$$

(b) *Pour tout $g > 0, h > 0$, il existe deux constantes H_1 et H_2 telles que*

$$0 < H_1 \leq F^{-1} \leq H_2 < \infty.$$

Une preuve du lemme 2 est donnée à l'annexe A.

Théorème 2 Soit $F_{g,h}(\cdot)$ la fonction de répartition de la variable aléatoire Bêta(g, h) et $F_{g,h}^{-1}(\cdot)$, son inverse. Soit

$$U, V \underset{\text{ind}}{\sim} \text{Uniforme}(0, 1),$$

et soit

$$X = F_{g,h}^{-1} \{ U F_{g,h}(d) + (1 - U) F_{g,h}(c) \}.$$

Si pour deux nombres réels $a, b > 1$,

$$V \leq \frac{1}{1 - c} \frac{(d - c)^{a+b-2}}{X - c} \left(\frac{\delta}{X - c} \right)^{a-1} \left(\frac{1 - \delta}{d - X} \right)^{b-1},$$

où $\delta = (a - 1) / (a + b - 2)$, alors X a la densité $f(x) =$

$$Af_1(x)f_2(x).$$

Une preuve du théorème 2 est donnée à l'annexe A.

Le théorème 1 nous donne l'algorithme suivant pour le tirage des échantillons de $f(\pi) \propto \pi^{s-1} (1 - \pi)^{h-1} (\pi - c)^{a-1} (d - \pi)^{b-1}$, $c < \pi < d, g, h > 0, a, b > 1$.

exactement parce que $\pi_i = (\theta - \sum_{j=1}^{\ell-1} \omega_j \pi_j) / \omega_i$, $\theta - \omega_i \leq \sum_{j=1}^{\ell-1} \omega_j \pi_j \leq \theta$. Autrement dit, la densité a posteriori conjointe n'est pas une fonction de π_i , et l'inférence a posteriori au sujet de $\pi_i = (\theta - \sum_{j=1}^{\ell-1} \omega_j \pi_j) / \omega_i$ découle de l'identité. Donc, il n'y a absolument aucune différence entre θ et $\sum_{j=1}^{\ell} \omega_j \pi_j$.

Théorème 1 *Sous le modèle resreint, la densité a posteriori conjointe, $p(\pi^{(i)}; \mu, \tau, \theta, s, \phi = 0)$, est appropriée.*

Une preuve du théorème 1 est donnée à l'annexe A.

Nous notons la différence entre les densités pour le modèle non restreint donné par (7) et le modèle restreint

donné par (9). Essentiellement, le terme

$$\left(\theta - \sum_{i=1}^{\ell-1} \omega_i \pi_i \right)_{s_i + \mu \tau} \times \left(1 - \frac{\theta}{\sum_{i=1}^{\ell-1} \omega_i \pi_i} \right)_{f_i + (1-\mu)\tau-1}$$

$$\times \theta^{t_0-1} (1-\theta)^{(1-t_0)\tau-1}$$

remplace $\pi_{s_i + \mu \tau}^{\ell} (1 - \pi_i)^{f_i + (1-\mu)\tau-1}$ dans (7).

Notons que, dans (9),

$$\pi_i = \frac{\theta - \sum_{j=1}^{\ell-1} \omega_j \pi_j}{\omega_i}.$$

Soit $a_i = s_i + \mu \tau$, $b_i = f_i + (1 - \mu) \tau$, $i = 1, \dots, \ell$. En

$$c_i = \frac{\omega_i}{\theta - \sum_{j=1, j \neq i}^{\ell-1} \omega_j \pi_j - \omega_i}$$

et

$$d_i = \frac{\omega_i}{\theta - \sum_{j=1, j \neq i}^{\ell-1} \omega_j \pi_j}, i = 1, \dots, \ell - 1.$$

Alors,

$$p(\pi_i | \pi^{(i)}; \mu, \tau, \theta, s, \phi = 0)$$

$$\propto \pi_i^{a_i-1} (1 - \pi_i)^{b_i-1} (d_i - c_i)^{c_i} (\pi_i - \pi_i)^{c_i} \quad (10)$$

$c_i > \pi_i$, $i = 1, \dots, \ell - 1$. Notons que cette fonction de densité comprend deux termes $\pi_i^{a_i-1} (1 - \pi_i)^{b_i-1}$ et $(\pi_i - c_i)^{c_i}$; notons l'échange entre a_i et b_i dans le deuxième terme. Le premier terme est la densité a posteriori conditionnelle sous le modèle non restreint, et le deuxième terme est une densité bêta généralisée [c'est-à-dire une loi bêta (b_i, a_i) dans l'intervalle (c_i, d_i)]. Donc, la densité bêta non restreinte est ajustée par la densité bêta généralisée. Dans le reste du document, nous désignons par GenBêta(a, b, c, d) la variable aléatoire bêta généralisée avec la fonction de densité,

$$p(x) = (x - c)^{a-1} (d - x)^{b-1} \{ (d - c)^{a+b-1} B(a, b) \}^{-1}$$

$$c \leq x \leq d, a > 1, b > 1.$$

et

$$E_{\pi}(\pi_i | \pi^{(i)}; \mu, \tau, \theta, s, \phi = 0) \approx c_i + (d_i - c_i) E_{\pi}(\pi_i | \mu, \tau, s)$$

Il découle de (13) que

$$\pi_i | \pi^{(i)}; \mu, \tau, \theta, s, \phi = 0 \sim \text{GenBêta}(a_i, b_i, c_i, d_i). \quad (13)$$

de (10) est

Donc, en combinant (11) et (12), notre approximation finale

$$(\pi_i - c_i)^{b_i} (d_i - \pi_i)^{a_i-1} \approx E[(\pi_i - c_i)^{b_i-1} (d_i - \pi_i)^{a_i-1}] \quad (12)$$

conduite approximation est

$(d_i - \pi_i)^{a_i-1}$ et sa variance soient petits. Alors, notre se-par conséquent, nous nous attendons à ce que $(\pi_i - c_i)^{b_i}$ par construction a_i et b_i sont relativement grands et, par conséquent, biais de $E[(\pi_i - c_i)^{b_i-1} (d_i - \pi_i)^{a_i-1}]$. En outre, cette dernière densité, $(\pi_i - c_i)^{b_i-1} (d_i - \pi_i)^{a_i-1}$ est un estimateur sans biais de $E[(\pi_i - c_i)^{b_i-1} (d_i - \pi_i)^{a_i-1}]$. Mais sous $\pi_i \sim \text{GenBêta}(a_i, b_i, c_i, d_i)$, $i = 1, \dots, \ell - 1$. où l'espérance est prise sur la distribution bêta généralisée

$$\times \frac{(\pi_i - c_i)^{b_i-1} (d_i - \pi_i)^{a_i-1}}{E[(\pi_i - c_i)^{b_i-1} (d_i - \pi_i)^{a_i-1}]}, c_i > \pi_i > d_i, \quad (11)$$

$$= \frac{(\pi_i - c_i)^{a_i-1} (d_i - \pi_i)^{b_i}}{B(a_i, b_i)}$$

$$= \frac{\int_{c_i}^{d_i} (\pi_i - c_i)^{a_i-1} (d_i - \pi_i)^{b_i} (\pi_i - c_i)^{b_i-1} (d_i - \pi_i)^{a_i-1} d\pi_i}{(\pi_i - c_i)^{a_i-1} (d_i - \pi_i)^{b_i-1} (\pi_i - c_i)^{b_i-1} (d_i - \pi_i)^{a_i-1}}$$

$$p_a(\pi_i | \pi^{(i)}; \mu, \tau, \theta, s, \phi = 0)$$

$$p_a(\pi_i | \pi^{(i)}; \mu, \tau, \theta, s, \phi = 0), \text{ nous avons}$$

Alors, en intégrant la constante de normalisation dans

$$c_i > \pi_i > d_i.$$

$$\propto (\pi_i - c_i)^{a_i-1} (d_i - \pi_i)^{b_i-1} (\pi_i - c_i)^{b_i-1} (d_i - \pi_i)^{a_i-1}$$

$$p_a(\pi_i | \pi^{(i)}; \mu, \tau, \theta, s, \phi = 0)$$

hypothèses, nous pouvons approximer (10) par

diffèrent de 0 et que d_i soit fort différent de 1. Sous cette assez faible, nous ne nous attendons pas à ce que c_i soit fort faibles. Premièrement, comme la contrainte étudiée est d'étudier (10) plus en profondeur en faisant deux approximations. Afin d'expliquer le gain de précision, nous essayons une théorie que des calculs.

grand). Cela est commode et avantageux tant du point de domaines par ordre de dénombrement (du plus petit au plus grand). Il convient de mentionner que nous avons placé les

quement si $X \sim \text{GenBêta}(a, b, c, d)$.

Autrement dit, $(X - c) / (d - c) \sim \text{Bêta}(a, b)$ si et uni-

Nous supposons que μ, τ, θ sont indépendants a priori avec $p(\mu, \tau, \theta) = p_1(\mu, \tau)p_2(\theta)$, où

$$p_1(\mu, \tau) = \frac{1}{1 + \tau^2}, \quad 0 < \mu < 1, \quad \tau \geq 0$$

comme dans (3), et $p_2(\theta)$ est donné par

$$\theta \sim \text{Bêta} \{\mu\tau_0, (1 - \mu\tau_0)\tau_0\}. \quad (6)$$

Pour le modèle restreint, nous considérons deux scénarios. En laissant $\tau_0 \rightarrow \infty$, θ devient un point matériel à μ_0 ,

et dans ce cas $\theta = \mu_0$ doit être spécifié par un praticien ; nous appellerons le modèle ajusté le modèle fixe (FI) ou modèle 2. Nous avons un deuxième scénario dans lequel un praticien spécifie μ_0 et τ_0 mais non θ ; nous appellerons ce modèle ajusté le modèle informatif (IN) ou modèle 3. Donc, il y a trois modèles, y compris le modèle non restreint. Afin de fournir un cadre unifié, il faut que tous les priors soient appropriés. La valeur exacte de θ est vraisemblablement inconnue dans la plupart des applications, ce qui peut donner lieu à des estimations n'ayant pas de cohérence interne.

Il convient de souligner que nous avons considéré un modèle supplémentaire pour faciliter l'étude du gain de précision de IN comparativement à FI. Pour les besoins de comparaison, nous voulons imposer un prior approprié mais non informatif à θ , de sorte que $\theta \sim \text{Uniforme}(0, 1)$ n'est pas un choix déraisonnable. En posant que $\mu_0 = 1/2$, $\tau_0 = 2$, nous obtenons $\theta \sim \text{Uniforme}(0, 1)$ avec ce prior, et nous appellerons le modèle ajusté le modèle uniforme (UN) ou modèle 4 ; naturellement, nous ne devons pas spécifier μ_0 et τ_0 . Il convient de souligner que le prior correspond à $\tau \rightarrow \infty$ est inapproprié car il correspond à $\theta \sim \text{Bêta}(0, 0)$. Nous ne considérons pas davantage ce modèle ; cependant, même si UN n'a pas de contrainte, nous le considérerons brièvement tout au long de l'exposé.

2.2 Inférence a posteriori

Nous envisageons de faire une inférence a posteriori au sujet de $\pi_i, i = 1, \dots, \ell$. Soit $\pi = (\pi_1, \dots, \pi_\ell)'$ et $\pi^{(i)} = (\pi_1, \dots, \pi_{i-1}, \pi_{i+1}, \dots, \pi_\ell)'$ [par exemple, $\pi^{(i)} = (\pi_1, \dots, \pi_{i-1})'$ tel que défini plus haut]. Nous utilisons le théorème de Bayes pour trouver les densités a posteriori conjointes de tous les paramètres. Premièrement, sous le modèle non restreint spécifié par (1), $\mathcal{G}(\pi, \mu, \tau | s) \propto \prod_{i=1}^{\ell} \frac{B\{s_i' + \mu\tau, f_i' + (1 - \mu)\tau\}}{\pi_{s_i' + \mu\tau - 1}^{f_i' + (1 - \mu)\tau - 1} (1 - \pi_i^{f_i' + (1 - \mu)\tau - 1})} \times \prod_{i=1}^{\ell} \frac{B\{\mu\tau, (1 - \mu)\tau\}}{B\{s_i' + \mu\tau, f_i' + (1 - \mu)\tau\}} \times \frac{1}{1 + \tau^2}.$ (7)

$0 < \pi_i < 1, i = 1, \dots, \ell, 0 < \mu < 1, \tau > 0, \theta = \mu - \omega_i \leq \sum_{j=1}^{\ell-1} \omega_j \pi_j \leq \sum_{j=1}^{\ell-1} \omega_j \pi_j' / \omega_\ell$. Il convient de souligner que la densité a posteriori conjointe (9) intègre la contrainte, $\sum_{i=1}^{\ell} \omega_i \pi_i' = \theta$,

(9)

$$p(\pi^{(i)}, \mu, \tau, \theta | s, \phi = 0) \propto \prod_{i=1}^{\ell} \frac{B\{s_i' + \mu\tau, f_i' + (1 - \mu)\tau\}}{\pi_{s_i' + \mu\tau - 1}^{f_i' + (1 - \mu)\tau - 1} (1 - \pi_i^{f_i' + (1 - \mu)\tau - 1})} \times \frac{B\{s_\ell' + \mu\tau, f_\ell' + (1 - \mu)\tau\}}{B\{s_i' + \mu\tau, f_i' + (1 - \mu)\tau\}} \times \left[\frac{\omega_\ell}{\theta - \sum_{i=1}^{\ell-1} \omega_i \pi_i'} \right] \left[\frac{\omega_\ell}{1 - \sum_{i=1}^{\ell-1} \omega_i \pi_i'} \right] \times \frac{1}{1 + \tau^2},$$

Nous obtenons la densité a posteriori conjointe pertinente en intégrant la contrainte ($\phi = 0$) dans (8). C'est-à-dire, $p(\pi^{(i)}, \mu, \tau, \theta | s, \phi = 0) \propto p(\pi^{(i)}, \mu, \tau, \theta, \phi = 0 | s)$, où

$$p(\pi^{(i)}, \mu, \tau, \theta, \phi = 0 | s) \propto \prod_{i=1}^{\ell} \frac{B\{s_i' + \mu\tau, f_i' + (1 - \mu)\tau\}}{\pi_{s_i' + \mu\tau - 1}^{f_i' + (1 - \mu)\tau - 1} (1 - \pi_i^{f_i' + (1 - \mu)\tau - 1})} \times \frac{1}{1 + \tau^2},$$

(8)

$$p(\pi^{(i)}, \mu, \tau, \theta, \phi | s) \propto \prod_{i=1}^{\ell} \frac{B\{s_i' + \mu\tau, f_i' + (1 - \mu)\tau\}}{\pi_{s_i' + \mu\tau - 1}^{f_i' + (1 - \mu)\tau - 1} (1 - \pi_i^{f_i' + (1 - \mu)\tau - 1})} \times \frac{B\{s_\ell' + \mu\tau, f_\ell' + (1 - \mu)\tau\}}{B\{s_i' + \mu\tau, f_i' + (1 - \mu)\tau\}} \times \left[\frac{\omega_\ell}{\phi + \theta - \sum_{i=1}^{\ell-1} \omega_i \pi_i'} \right] \left[\frac{\omega_\ell}{\phi + \theta - \sum_{i=1}^{\ell-1} \omega_i \pi_i'} \right] \times \frac{1}{1 + \tau^2},$$

de $\pi^{(i)}, \mu, \tau, \theta, \phi$ est

Sous le modèle restreint, la densité a posteriori conjointe

Une preuve du lemme 1 est donnée à l'annexe A.

conjointe, $\mathcal{G}(\pi, \mu, \tau | s)$, est appropriée.

Lemme 1 Sous le modèle non restreint la densité a posteriori

$$0 < \pi_i < 1, 0 < \mu < 1, \tau > 0, i = 1, \dots, \ell.$$

loi a posteriori bêta. Dans l'un et l'autre cas, notre procédure peut être appliquée.

Le plan du présent article est le suivant. À la section 2,

nous décrivons la méthodologie. En particulier, nous décrivons le modèle bêta-binomial standard et nous élaborons deux autres modèles pour intégrer l'information supplémentaire en utilisant les lois a priori. Nous décrivons aussi l'inférence a posteriori et la façon d'effectuer les calculs non standard. À la section 3, nous décrivons un exemple portant sur l'obésité, et une étude en simulation pour évaluer empiriquement les propriétés statistiques de nos modèles. À la section 4, nous présentons nos conclusions. Nous discutons aussi de la façon de faire une inférence prédictive bayésienne pour les proportions de population finie. Bien que nous discussions de données binaires, nous montrons aussi comment notre méthode peut être étendue à des données polychotomiques.

2. Méthodologie

Nous montrons comment intégrer la contrainte dans le modèle bêta-binomial de deux façons, ce qui fournit un ensemble de modèles de rechange. À la section 2.1, nous décrivons les modèles et à la section 2.2, nous décrivons l'inférence sur les estimations des probabilités en utilisant une approximation. À la section 2.3, nous décrivons les calculs et un nouvel algorithme.

2.1 Modèles

Nous supposons que des données binaires sont disponibles en provenance de ℓ petits domaines, et nous supposons que la probabilité qu'un individu réponde dans le i^{e} domaine est π_i , $i = 1, \dots, \ell$. Soit n_i le nombre d'individus échantillonnés dans le i^{e} domaine, $i = 1, \dots, \ell$. Soit aussi s_i le nombre d'individus possédant la caractéristique et $f_i = n_i - s_i$ le nombre d'individus sans la caractéristique dans le i^{e} domaine, $i = 1, 2, \dots, \ell$. Alors, le modèle hiérarchique bayésien bêta-binomial standard est

$$s_i | \pi_i \sim \text{Binomiale}(n_i, \pi_i), \quad (1)$$

$$\pi_i | \mu, \tau \sim \text{Bêta}(\mu\tau, (1-\mu)\tau), \quad i = 1, \dots, \ell \quad (2)$$

$$p(\mu, \tau) = \frac{1}{(1+\tau)^2}, \quad 0 < \mu < 1, \tau \geq 0. \quad (3)$$

Nous utilisons un prior de rétrécissement pour τ , parce qu'il est approprié et non informatif et qu'il n'y a pas de priors conjugués. Les priors de la forme $p(\tau) \propto 1/\tau$ sont déconseillés; voir, par exemple, Gelman (2006). D'autres options sont les demi-densités de Cauchy et les densités

gamma (il serait nécessaire de spécifier les hyper-paramètres). Donc, nous appellerons le modèle spécifié par (1), (2) et (3) le modèle non restreint (NR) ou modèle 1.

Ensuite, nous décrivons le modèle restreint, qui est une extension du modèle non restreint. Nous obtenons une simple combinaison linéaire des probabilités binomiales. En posant que $\pi_i = s_i/n_i$ et

$$\omega_i = \frac{\sum_{i'} n_{i'}}{n_i}, \quad i = 1, \dots, \ell,$$

nous avons

$$\frac{\sum_{i'} n_{i'}}{\sum_{i'} s_{i'}} = \sum_{i'} \omega_i \pi_{i'}$$

Donc, en prenant les π_i inconnues, la combinaison linéaire est $\sum_{i'} \omega_i \pi_{i'}$.

Par conséquent, nous devons faire un ajustement dans (2) afin d'incorporer la contrainte $\sum_{i'} \omega_i \pi_{i'} = \theta$ conditionnellement à θ . Nous le faisons en introduisant la variable $\phi = \sum_{i'} \omega_i \pi_{i'} - \theta$; de sorte que la contrainte est équivalente à $\phi = 0$. Maintenant, une des variables, π_i , $i = 1, \dots, \ell$, est redondante. Il convient de souligner que l'on peut choisir n'importe laquelle des π_1, \dots, π_ℓ , et, sans perte de généralité, pour faciliter l'exposé, nous choisissons π_ℓ . Donc, pour intégrer la contrainte, nous transformons π_ℓ en $\phi = \sum_{i'} \omega_i \pi_{i'} - \theta$, en gardant les $\pi_1, \dots, \pi_{\ell-1}$ non transformées, et nous posons que $\pi_{(\ell)} = (\pi_1, \dots, \pi_{\ell-1})'$. Comme le jacobien est $1/\omega_\ell$

$$p(\pi_{(\ell)}, \phi | \mu, \tau, \theta) =$$

$$\frac{1}{\prod_{i=1}^{\ell-1} \pi_i^{\mu\tau-1} (1-\pi_i)^{(1-\mu)\tau-1}} B\{\mu\tau, (1-\mu)\tau\} \times \left[\frac{\omega_\ell}{\sum_{i=1}^{\ell-1} \omega_i \pi_i} \phi + \theta - 1 \right]^{1-\mu\tau-1} \left[\frac{\omega_\ell}{\sum_{i=1}^{\ell-1} \omega_i \pi_i} \phi + \theta - \omega_\ell \right]^{(1-\mu)\tau-1} B\{\mu\tau, (1-\mu)\tau\} \quad (4)$$

où

$$0 < \pi_i < 1, \quad i = 1, \dots, \ell,$$

$$0 < \mu < 1, \tau > 0, \phi + \theta - \omega_\ell \leq \sum_{i=1}^{\ell-1} \omega_i \pi_i \leq \phi + \theta,$$

et

$$\pi_{\ell'} = \frac{\omega_{\ell'}}{\sum_{i=1}^{\ell-1} \omega_i \pi_i + \phi + \theta}. \quad (5)$$

Notons que la densité a priori conjointe de $(\pi_{(\ell)}, \phi)$ dans (4) est bien définie. Nous souhaitons prendre $\phi = 0$ dans (5) pour intégrer la contrainte, mais quand $\phi = 0$, la densité conjointe de $\pi_{(\ell)}$ n'est pas bien définie.

répartition proportionnelle, ils peuvent être proportionnels aux tailles d'échantillon. Le but de l'intégration d'information a priori au sujet des probabilités binomiales est d'accroître la précision et, en même temps, il faut contrôler le biais.

Il est nettement plus facile pour un praticien d'enquête de spécifier la valeur de la probabilité globale que celle des probabilités des domaines individuels. Autrement dit, la probabilité globale peut être spécifiée avec relativement moins d'erreur que les probabilités individuelles. Naturellement, on peut spécifier la probabilité globale en utilisant de l'information a priori (une enquête antérieure, un recensement ou des dossiers administratifs) et la spécification de la probabilité globale dépendra donc de la qualité de l'information a priori. D'où le problème relève naturellement du paradigme bayésien, parce que nous intégrons l'information a priori au sujet d'un paramètre au moyen d'une distribution. Donc, il y aura un gain de précision à cause de l'information supplémentaire. Cependant, un praticien peut encore procéder en l'absence d'information a priori. On peut utiliser le ratio du succès total et de la taille totale d'échantillon sur les domaines pour former une spécification raisonnable de la probabilité globale qui n'est habituellement pas la probabilité d'intérêt. Cette estimation aura une beaucoup plus grande précision que celle pour les domaines individuels. Il y aura encore un gain de précision, mais de toute évidence, ce gain est dû à l'utilisation de données courantes (utilisation double) et de la contrainte.

Un exemple d'enquête dans laquelle de l'information fiable peut être obtenue pour procéder à l'étalonnage est la National Health Interview Survey (NHIS) qui est réalisée annuellement par le National Center for Health Statistics pour évaluer un aspect de la santé de la population américaine. Il s'agit d'une enquête représentative de la population qui porte sur de nombreux indicateurs de la santé ; l'un de ces indicateurs est le nombre de visites chez le médecin au cours des deux dernières semaines, et une entité informative est la proportion de personnes qui ont rendu au moins une visite à un médecin l'année précédente (par exemple, Nandram et Choi 2002). Ces proportions sont utiles pour les petits domaines formés par croisement de l'âge, de la race et du sexe pour un État particulier l'année précédente. Comme les estimations à l'échelle d'un État varient fort peu d'une année à l'autre, l'estimation globale pour l'année qui précède immédiatement la dernière année peut être utilisée comme valeur de référence fiable pour la dernière année. S'il n'est pas possible d'obtenir une estimation fiable pour l'étalonnage, on peut constituer une loi a priori informative. Par exemple, on peut utiliser la méthode des moments pour égaliser la moyenne d'échantillon et la variance d'échantillon des estimations globales pour les quelques années passées à la moyenne et à la variance d'une loi bêta pour obtenir une

Polyà à une loi prédictive des paramètres d'une population

finie.

Notre procédure est reliée à l'étalonnage externe qui a lieu quand un estimateur préspecifié est obtenu de sources externes, telles qu'une enquête différente, un recensement ou d'autres dossiers administratifs. Dans l'étalonnage, on veut que la somme des parties soit égale au tout. Par exemple, quand des enquêtes sont réalisées au cours du temps, il existe habituellement des enquêtes mensuelles et des enquêtes annuelles qui sont de nettement meilleure qualité que les enquêtes mensuelles. Quand des estimations sont produites d'après les enquêtes mensuelles de telle façon que la somme de ces estimations concorde avec les totaux des enquêtes annuelles, il existe une protection contre l'échec du modèle et, par conséquent, une amélioration des estimations (c'est-à-dire biais réduit et éventuellement augmentation de la précision). Ces problèmes sont fréquents dans les organismes gouvernementaux, spécialement en ce qui concerne l'emploi et les ventes ; voir Hillmer et Trabelsi (1987) pour un exemple portant sur les ventes au détail des quincailleries provenant du U.S. Census Bureau.

L'information a priori issue de l'étalonnage externe améliore la précision, mais peut également produire des estimateurs gravement biaisés. Cela dépendra de la mesure dans laquelle l'enquête courante diffère des précédentes. Nandram, Toto et Choi (2011) ont appliqué l'étalonnage externe pour estimer les valeurs moyennes de petits domaines en population finie. La contrainte est la moyenne de population pour l'ensemble de la population, qui est une valeur préspecifiée pouvant, de nouveau, être tirée d'une enquête antérieure, d'un recensement ou de dossiers administratifs. Dans nos travaux courants, nous n'intégrons pas l'information au sujet d'une combinaison linéaire des valeurs de population finie, mais nous incorporons plutôt l'information au sujet d'une combinaison linéaire des paramètres de superpopulation (ici les probabilités binomiales).

Nous considérons le problème dans lequel les dénombrements binomiaux sont obtenus auprès de petits domaines similaires, et où l'inférence est requise au sujet des probabilités binomiales. Dans la conclusion, nous discutons de la façon d'étendre notre méthode pour obtenir la loi prédictive des proportions de population finie. Le modèle bêta-binomial standard peut s'avérer inadéquat et l'information a priori supplémentaire peut être incorporée. Nous postulons qu'il y a une augmentation de la précision par rapport au modèle bêta-binomial standard pour petits domaines quand l'information a priori au sujet de la moyenne pondérée des probabilités (par exemple, moyenne des probabilités) est incorporée. Autrement dit, nous intégrons l'information a priori au sujet d'une combinaison linéaire de probabilités (une moyenne pondérée). Les poids peuvent être proportionnels aux tailles de population et, sous une

Une analyse bayésienne des probabilités de réponse dans les petits domaines sous une contrainte

Balagobin Nandram et Hasanjan Sayit¹

Résumé

De nombreuses enquêtes par sondage comprennent des questions suscitant une réponse binaire (par exemple, obèse, non obèse) pour un certain nombre de petits domaines. Une inférence est requise au sujet de la probabilité d'une réponse positive (par exemple obèse) dans chaque domaine, la probabilité étant la même pour tous les individus dans chaque domaine et différente entre les domaines. Étant donné le peu de données dans les domaines, les estimateurs directs ne sont pas fiables et il est nécessaire d'utiliser des données provenant d'autres domaines pour améliorer l'inférence pour un domaine particulier. Essentiellement, il est supposé a priori que les domaines sont similaires, si bien que le choix d'un modèle hiérarchique bayésien, le modèle beta-binomial standard, est naturel. L'innovation tient au fait qu'un praticien peut disposer d'information a priori supplémentaire au sujet d'une combinaison linéaire des probabilités. Par exemple, une moyenne pondérée des probabilités est un paramètre, et l'information peut être obtenue au sujet de ce paramètre, ce qui rend le paradigme bayésien approprié. Nous avons modifié le modèle beta-binomial standard pour petits domaines afin d'y intégrer l'information a priori sur la combinaison linéaire des probabilités, que nous appelons une contrainte. Donc, il existe trois cas. Le praticien a) ne spécifie pas de contrainte, b) spécifie une contrainte et le paramètre entièrement et c) spécifie une contrainte et l'information qui peut être utilisée pour construire une loi a priori pour le paramètre. L'échantillonneur de Gibbs « gridly » est utilisé pour ajuster les modèles. Pour illustrer notre méthode, nous prenons l'exemple de l'obésité chez les enfants dans la National Health and Nutrition Examination Survey dans laquelle les petits domaines sont formés par croisement de l'école (cycle secondaire inférieur ou supérieur), de l'ethnicité (blanche, noire, mexicaine) et du sexe (masculin, féminin). Nous procédons à une étude en simulation pour évaluer certaines caractéristiques statistiques de notre méthode. Nous avons montré que le gain de précision au-delà de (a) est dans l'ordre où (b) est plus grand que (c).

Mots clés : Algorithme d'acceptation-rejet ; loi binomiale ; loi beta généralisée ; échantillonneur de Gibbs « gridly » ; simulation.

1. Introduction

L'utilisation de modèles pour « emprunter de l'information » dans l'estimation sur petits domaines est une pratique standard (Rao 2003). Étant donné le peu de données pré-sentes dans chaque domaine, les estimations directes pour les petits domaines ne sont habituellement pas fiables. Notre procédure permet à un praticien d'intégrer l'information a priori au sujet d'une combinaison linéaire de probabilités binomiales, une pour chaque domaine. Il s'agit d'une contrainte que nous incluons sous forme de moyenne pondérée des probabilités de domaine dans le modèle beta-binomial standard. La moyenne pondérée peut être supposée connue ou inconnue. Dans le cas où cette valeur est inconnue, nous considérons le scénario où il existe une certaine information qui peut être obtenue auprès d'un expert sous forme de loi a posteriori. Cette situation diffère de la pratique courante en échantillonnage fondé sur le plan de sondage dans lequel l'information auxiliaire est intégrée, comme dans les estimateurs par le ratio et par la régression (Cochran 1977). Quand la valeur peut être spécifiée exactement, la précision augmente, parce que l'information a priori est intégrée dans le domaine.

Le modèle beta-binomial a été étudié extensivement. Par exemple, Nandram et Sedransk (1993), Nandram (1998) et Nandram et Choi (2002) montrent comment faire une inférence prédictive bayésienne des proportions de petit domaine en population finie pour des données binomiales et multinomiales. Ces modèles supposent que les probabilités binomiales ont un effet en commun, ce qui permet le groupement adaptatif des données provenant des petits domaines (ou grappes). Cependant, il est possible d'améliorer encore davantage ces modèles en incluant de l'information supplémentaire en se servant de covariables par la voie de modèles linéaires généralisés (par exemple, voir Ghosh, Natarajan, Stroud et Carlin 1998). Il convient de mentionner que dans aucun de ces travaux il n'est proposé des moyens d'intégrer l'information a priori au sujet de la combinaison linéaire de paramètres du modèle. Des gains de précision importants sont attendus quand ce genre d'information a priori est intégré dans les modèles pour petits domaines ; voir Silvapulle et Sen (2006) pour une longue discussion de l'inférence statistique sous contrainte. Il convient aussi de souligner que Lazar, Meeden et Nelson (2008) ont montré comment inclure des contraintes dans une approche bayésienne non paramétrique au moyen d'un schéma d'urne de

Scott, A.J., et Wild, C.J. (2009). Population-based case-control studies. Dans *Handbook of Statistics 29B; Sample Surveys: Inference and Analysis*, (Eds., D. Pfeffermann et C.R. Rao). Amsterdam : Hollande du Nord, 431-453.

Skinner, C.J., Holt, D. et Smith, T.M.F. (Eds.) (1989). *Analysis of complex surveys*. New York : John Wiley & Sons, Inc.

Skinner, C.J. (1994). Sample models and weights. *Proceedings of the Section on Survey Research Methods*, 133-142.

Smith, T.M.F. (1988). To weight or not to weight, that is the question. Dans *Bayesian Statistics 3*, (Eds., J.M. Bernardo, M.H. Degroot, D.V. Lindley et A.F.M. Smith), Oxford University Press, 437-451.

Sugden, R.A., et Smith, T.M.F. (1984). Ignorable and informative designs in survey sampling inference. *Biometrika*, 71, 495-506.

Sverchkov, M., et Pfeffermann, D. (2004). Prévion des totaux de population finie basée sur la distribution échantillonale. *Techniques d'enquête*, 30, 87-101.

Wu, Y.Y., et Fuller, W.A. (2006). Estimation of regression coefficients with unequal probability samples. *Proceedings of the American Statistical Association, Section on Survey Research Methods*, 3892-3899.

Samdal, C.-E., et Wright, R. (1984). Cosmetic form of estimators in survey sampling. *Scandinavian Journal of Statistics*, 11, 146-156.

Pfeffermann, D., Moura, F.A.S. et Nascimento-Silva, P.L. (2006). Multilevel modeling under informative sampling. *Biometrika*, 93, 943-959.

Pfeffermann, D., et Sverchkov, M. (2007). Small area estimation under informative probability sampling of areas and within the selected areas. *Journal of the American Statistical Association*, 102, 1427-1439.

Pfeffermann, D., et Sverchkov, M. (2009). Inference under Informative Sampling. Dans *Handbook of Statistics 29B; Sample Surveys: Inference and Analysis*, (Eds., D. Pfeffermann et C.R. Rao). Amsterdam : Hollande du Nord, 455-487.

Pfeffermann, D., et Landsman, V. (2011). Are private schools better than public schools? Appraisal for Ireland by methods for observational studies. *The Annals of Applied Statistics*, 5, 1726-1751.

Pfeffermann, D., et Sikov, N. (2011). Imputation and estimation under nonignorable nonresponse in household surveys with missing covariate information. *Journal of Official Statistics*, 27, 181-209.

Rubin, D.B. (1976). Inference and missing data. *Biometrika*, 63, 605-614.

Rubin, D.B. (1985). The use of propensity scores in applied Bayesian inference. Dans *Bayesian Statistics 2*, (Eds., J.M. Bernardo, M.H. Degroot, D.V. Lindley et A.F.M. Smith), Elsevier Science Publishers B.V., 463-472.

- Chang, T., et Kott, P.S. (2008). Using calibration weighting to adjust for nonresponse under a plausible model. *Biometrika*, 95, 555-571.
- Chen, J., et Sitter, R.R. (1999). A pseudo empirical likelihood approach to the effective use of auxiliary information in complex surveys. *Statistica Sinica*, 9, 385-406.
- Chaudhuri, S., Handcock, M.S. et Rendall, M.S. (2010). A conditional empirical likelihood approach to combine sampling design and population level information. Rapport technique No. 3/2010, National University of Singapore, Singapore, 117546.
- DeMeets, D., et Halperin, M. (1977). Estimation of simple regression coefficients in samples arising from sub-sampling procedures. *Biometrics*, 33, 47-56.
- DuMouchel, W.H., et Duncan, G.L. (1983). Using sample survey weights in multiple regression analysis of stratified samples. *Journal of the American Statistical Association*, 78, 535-543.
- Feder, M. (2011). Fitting Regression Models to Complex Survey Data - Gelman's Estimator Revisited. Dans Proceedings of the ISI meeting, Ireland, (www.isi2011.ie).
- Francisco, C.A., et Fuller, W.A. (1991). Quantile estimation with a complex survey design. *The Annals of Statistics*, 19, 454-469.
- Fuller, W.A. (1975). Regression analysis for sample surveys. *Sankhyā, Series C*, 37, 117-132.
- Fuller, W.A. (2002). Estimation par régression appliquée à l'échantillonnage. *Techniques d'enquête*, 28, 5-25.
- Gelman, A., Carlin, J.B., Stern, H.S. et Rubin, D.B. (2003). *Bayesian Data Analysis*, deuxième édition. Londres : CRC Press.
- Gelman, A. (2007). Struggles with survey weighting and regression modeling (avec discussion). *Statistical Science*, 22, 153-164.
- Godambe, V.P., et Thompson, M.E. (1986). Parameters of superpopulation and survey population: their relationships and estimation. *Revue Internationale de Statistique*, 54, 127-138.
- Godambe, V.P., et Thompson, M.E. (2009). Estimating functions and survey sampling. Dans *Handbook of Statistics* 29B: *Sample Surveys: Inference and Analysis*, (Eds., D. Pfeffermann et C.R. Rao). Amsterdam : Hollande du Nord, 83-101.
- Goldstein, H. (1986). Multi-level mixed linear model analysis using iterative generalized least squares. *Biometrika*, 73, 43-56.
- Häjek, J. (1971). Comments on a paper by D. Basu. Dans *Foundations of Statistical Inference*, (Eds., V.P. Godambe et D.A. Sprott). Toronto : Holt, Rinehart and Winston.
- Hartley, H.O., et Rao, J.N.K. (1968). A new estimation theory for sample surveys. *Biometrika*, 55, 547-557.
- Holt, D., Smith, T.M.F. et Winter, P.D. (1980). Regression analysis of data from complex surveys. *Journal of the Royal Statistical Society, Series A*, 143, 474-487.
- Jewell, N.P. (1985). Least squares regression with data arising from stratified samples of the dependent variable. *Biometrika*, 72, 11-21.
- Kasprzyk, D., Duncan, G.J., Kaiton, G. et Singh, M.P. (1989, Eds.). *Panel Surveys*. New York : John Wiley & Sons, Inc.
- Kim, J.K. (2009). Calibration estimation using empirical likelihood in survey sampling. *Statistica sinica*, 19, 145-157.
- Kott, P.S. (2009). Calibration Weighting: Combining Probability Samples and Linear Prediction Models. Dans *Handbook of Statistics* 29B: *Sample Surveys: Inference and Analysis*, (Eds., D. Pfeffermann et C.R. Rao). Amsterdam : Hollande du Nord, 55-82.
- Krieger, A.M., et Pfeffermann D. (1997). Testing of distribution functions from complex sample surveys. *Journal of Official Statistics*, 13, 123-142.
- Little, R.J.A. (1982). Models for non-response in sample surveys. *Journal of the American Statistical Association*, 77, 237-249.
- Little, R.J.A. (2004). To model or not to model? Competing modes of inference for finite population sampling. *Journal of the American Statistical Association*, 99, 546-556.
- Magee, L. (1998). Improving survey-weighted least squares regression. *Journal of the Royal Statistical Society, Series B*, 60, 115-126.
- Nathan, G., et Holt, D. (1980). The effect of survey design on regression analysis. *Journal of the Royal Statistical Society, Series B*, 42, 377-386.
- Orchard, T., et Woodbury, M.A. (1972). A missing information principle: theory and application. *Proceedings of the 6th Berkeley Symposium on Mathematical Statistics and Probability*, 1, 697-715.
- Owen, A.B. (1988). Empirical likelihood ratio confidence intervals for a single functional. *Biometrika*, 75, 237-249.
- Owen, A.B. (2001). *Empirical likelihood*. New York : Chapman & Hall.
- Pfeffermann, D., et Holmes, D. (1985). Robustness consideration in the choice of method of inference for regression analysis of survey data. *Journal of the Royal Statistical Society, Series A*, 148, 268-278.
- Pfeffermann, D., et Smith, T.M.F. (1985). Regression models for grouped populations in cross-section surveys. *Revue Internationale de Statistique*, 53, 37-59.
- Pfeffermann, D. (1993). The role of sampling weights when modeling survey data. *Revue Internationale de Statistique*, 61, 317-337.
- Pfeffermann, D. (1996). The use of sampling weights for survey data analysis. *Statistical Methods in Medical Research*, 5, 239-261.
- Pfeffermann, D., Krieger, A.M. et Rinott, Y. (1998a). Parametric distributions of complex survey data under informative probability sampling. *Statistica Sinica*, 8, 1087-1114.
- Pfeffermann, D., Skinner, C.J., Holmes, D.J., Goldstein, H. et Rasbash, J. (1998b). Weighting for unequal selection probabilities in multi-level models (avec discussion). *Journal of the Royal Statistical Society, Series B*, 60, 23-76.
- Pfeffermann, D., et Sverchokov, M. (1999). Parametric and semi-parametric estimation of regression models fitted to survey data. *Sankhyā*, 61, 166-186.
- Pfeffermann, D., et Sverchokov, M. (2003). Fitting generalized linear models under informative probability sampling. Dans *Analysis of Survey Data*, (Eds., R.L. Chambers et C.J. Skinner). New York : John Wiley & Sons, Inc, 175-195.
- Statistique Canada, N° 12-001-X au catalogue

Les taux de couverture sont presque systématiquement inférieurs aux niveaux nominaux, mais, dans le cas des IC classiques, la sous-couverture est généralement inférieure à 4 %. Les deux exceptions s'observent dans le cas où les intervalles de confiance sont fondés sur les estimateurs par les moindres carrés (OLS) (forte sous-couverture) et sur l'estimateur du maximum de vraisemblance (MLE) de la pente (sous-couverture de 7 % au niveau nominal de 90 %), résultats qui expliquent le biais de ces estimateurs. Les pourcentages de sous-couverture lorsque l'on utilise la méthode bootstrap ordinaire sont généralement un peu plus importants, sauf dans le cas de l'IC pour l'ordonnée à l'origine fondé sur β_{sel} , dont la sous-couverture est plus prononcée.

Remarque 15. Nous avons également calculé l'IC classique d'après les erreurs-types estimées sous la distribution aléatoire (équation 4.4) et sous le modèle d'échantillon (équation 4.5), mais sauf dans le cas des estimateurs β^{pw} et β^q , la sous-couverture de ces intervalles était un peu plus importante que les taux de couverture dans la Figure 2, à cause de la sous-estimation des erreurs-types réelles par ces estimateurs de l'erreur-type comme il a été discuté plus haut. Le même phénomène a été observé en utilisant la « méthode du bootstrap studentisé » avec ces estimateurs de l'erreur-type, ce qui peut de nouveau être expliqué par la sous-estimation des erreurs-types réelles. L'utilisation d'un IC fondé sur une méthode du bootstrap plus avancée, telle que le bootstrap double, pourrait corriger cette sous couverture.

5. Conclusion

Dans le présent article, je discute de l'utilisation de diverses procédures proposées dans la littérature pour tenir compte de l'échantillonnage informatif et de la non-réponse de type NMAR sous modélisation des données d'enquête. L'étude empirique est limitée jusqu'à présent au cas de la régression linéaire et de l'échantillonnage à un degré, et une extension évidente consisterait à considérer d'autres modèles et l'échantillonnage par grappes. La présente étude illustre l'absence de biais ou la quasi-absence de biais de tous les estimateurs ponctuels pris en considération, mais les estimateurs de variance classiques sous-estiment les variances réelles dans la plupart des cas, puisqu'ils ne tiennent pas compte des opérations supplémentaires nécessaires pour calculer les estimateurs ponctuels correspondants. Les estimateurs bootstrap de la variance sont de nettement meilleurs estimateurs de variance dans ces conditions. Les intervalles de confiance appliqués dans la présente étude produisent une faible sous-couverture dans la plupart des cas, mais ils devraient être améliorés, éventuellement en

Remerciements

L'auteur est reconnaissant envers Moshe Feder, qui a exécuté l'étude empirique, formulé de nombreux commentaires et fait des suggestions utiles. L'auteur est également reconnaissant envers Pedro Silva, qui a fait des remarques constructives au sujet d'une ébauche antérieure de l'article, ainsi qu'à l'égard de trois examinateurs pour leur lecture attentive et leurs commentaires en dépit du délai très court qui leur avait été donné. La présente étude est financée par la bourse de recherche n° RES-062-23-2316 de l'ESRC du Royaume-Uni.

Bibliographie

Binder, D.A. (1983). On the variances of asymptotically normal estimators from complex surveys. *Revue Internationale de Statistique*, 51, 279-292.

Binder, D., et Roberts, G. (2009). Design and model based inference for model parameters. Dans *Handbook of Statistics 29B: Sample Surveys: Inference and Analysis*, (Eds., D. Pfeffermann et C.R. Rao). Amsterdam : Hollande du Nord, 33-54.

Breckling, J.U., Chambers, R.L., Dorfman, A.H., Tarn, S.M. et Welsh, A.H. (1994). Maximum likelihood inference from sample survey data. *Revue Internationale de Statistique*, 62, 349-363.

Brick, J.M., et Montaquila, J.M. (2009). Nonresponse and weighing. Dans *Handbook of Statistics 29A: Sample Surveys: Inference and Analysis*, (Eds., D. Pfeffermann et C.R. Rao). Amsterdam : Hollande du Nord, 163-185.

Chambers, R.L., Dorfman, A.H. et Wang, S. (1998). Limited information likelihood analysis of survey data. *Journal of the Royal Statistical Society, Séries B*, 60, 397-411.

Chambers, R.L., et Skinner, C.J. (2003, Eds.), *Analysis of survey data*. New York : John Wiley & Sons, Inc.

Chambliss, L.E., et Boyle, K.E. (1985). Maximum likelihood methods for complex sample data: Logistic, regression and discrete proportional hazards models. *Communication in Statistics-Theory and Methods*, 14, 1377-1392.

du choix des poids optimaux $a_i(\alpha)$, dans le cas de β_f , l'estimateur de variance ne tient pas compte de l'imputation aléatoire des vecteurs (y_i, x_i) pour $i \in U - s$, et dans le cas de β_{mle} et β_{sel} , les estimateurs de variance ne tiennent pas compte de l'estimation des probabilités $\Pr(i \in s | y_i, x_i)$. Cette sous-estimation de la variance est corrigée dans presque tous les cas par l'utilisation de la méthode du bootstrap, voir, en particulier, l'estimation des variances de β_f, β_{mle} et β_{sel} .

Tableau 1
Moyennes, erreurs-types (e-t.) et racines carrées des moyennes des estimations de variance. Modèle de population : $E_p(y_j) = 2 + 1 \times x_j$, $\text{Var}_p(y_j) = (1 + 0,2x_j)^2 V_j + 1$

Méthode	Ordonnée à l'origine- β_0					Pente- β_1				
	Moyenne	E.-T.	Rac. car. [moyenne (est. var.)]	M.E.	BS	Moyenne	E.-T.	Rac. car. [moyenne (est. var.)]	M.E.	BS
β_{ols}	2,251	0,133	0,135	0,139	0,140	1,046	0,048	0,048	0,049	0,049
β_f	2,006	0,133	0,126	0,126	0,135	0,999	0,051	0,041	0,041	0,052
β_{pw}	2,008	0,166	0,167	0,169	0,157	0,998	0,059	0,055	0,055	0,056
β_{mg}	2,017	0,158	0,154	0,156	0,154	0,995	0,056	0,050	0,050	0,055
β_q	2,011	0,153	0,157	0,159	0,147	0,999	0,054	0,051	0,051	0,052
β_{mg-y}	2,020	0,156	0,152	0,154	0,153	0,996	0,055	0,049	0,050	0,054
β_{mle}	1,960	0,159	----	0,143	0,152	1,026	0,054	----	0,046	0,053
β_{sel}	2,031	0,164	0,143	0,143	0,159	0,995	0,058	0,049	0,049	0,057

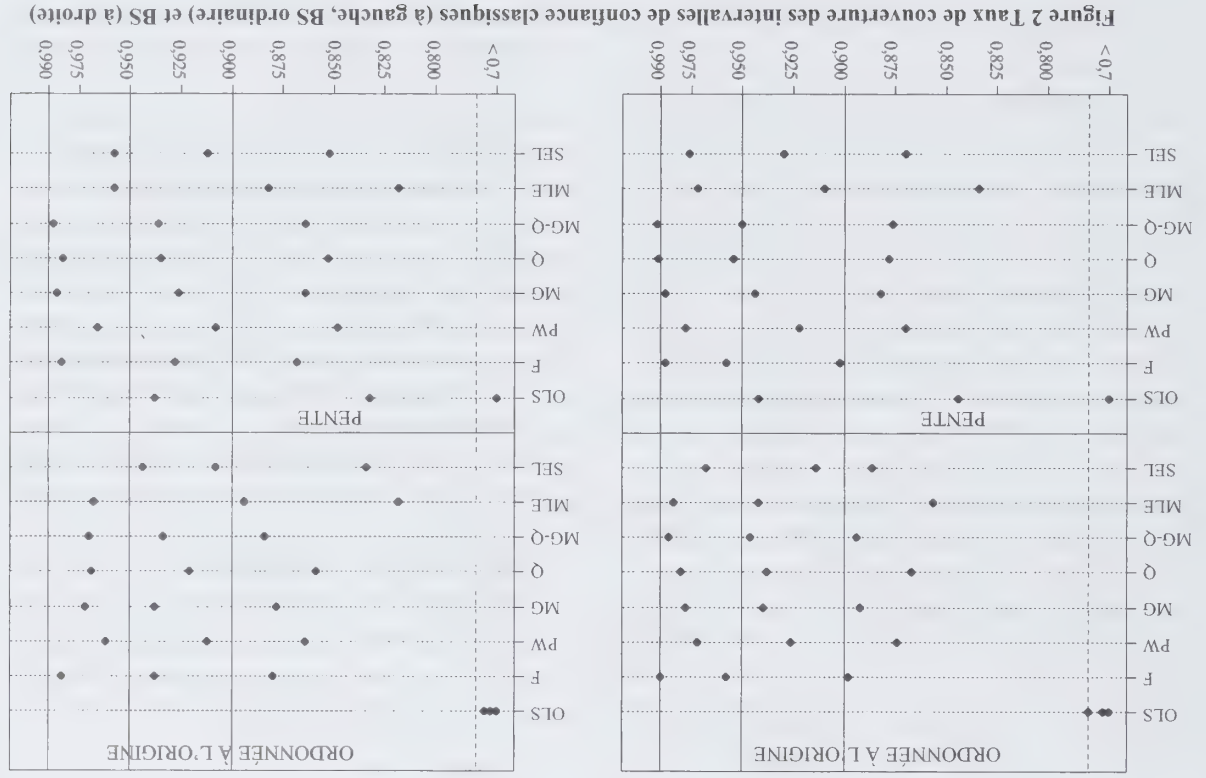


Figure 2 Taux de couverture des intervalles de confiance classiques (à gauche, BS ordinaire) et BS (à droite)

Comme prévu, étant donné l'utilisation d'un scénario d'échantillonnage informatif, l'estimateur par les moindres carrés ordinaires présente un biais relativement important de 12 % (5 %) quand on estime l'ordonnée à l'origine (la pente). Tous les autres estimateurs sont pratiquement sans biais, excepté l'estimateur du maximum de vraisemblance $\hat{\beta}_{\text{mle}}$ dont les biais sont de 2 % et de 1,5 %. La quasi-absence de biais de l'estimateur de la vraisemblance empirique $\hat{\beta}_{\text{sel}}$ est particulièrement encourageante, étant donné l'estimation non paramétrique un peu grossière des probabilités $\tau_i = \Pr(i \in s \mid y_i, x_i)$. Notons aussi que les E.-T. empiriques de cet estimateur sont semblables à celles de l'estimateur PW à pondération probabiliste. Le biais faible, mais statistiquement significatif, de $\hat{\beta}_{\text{mle}}$ s'explique par le fait que nous supposons que la distribution sous le modèle de population est normale, ce qui est incorrect, comme nous l'avons mentionné et illustré plus haut.

En ce qui concerne la précision, l'estimateur par les moindres carrés ordinaires est celui dont les erreurs-types sont les plus faibles, mais $\hat{\beta}_f$ donne presque les mêmes erreurs-types (et est sans biais). Cela tient au fait que cet estimateur utilise l'information supplémentaire sur la stratification qui n'est pas employée dans les autres estimateurs. Notons que $\hat{\beta}_{\text{mg}}$, $\hat{\beta}_{\text{mg-q}}$ et particulièrement $\hat{\beta}_q$ donnent de meilleurs résultats que $\hat{\beta}_{\text{pw}}$, mais que $\hat{\beta}_{\text{mg-q}}$ n'offre pas d'améliorations par rapport à $\hat{\beta}_q$.

Remarque 14. À la suite de la présentation de l'article au Symposium de Statistique Canada en 2011, Jean-François Beaumont a suggéré de remplacer les poids $\hat{\tau}_i^{-1}$ utilisés pour le calcul de $\hat{\beta}_{\text{sel}}$ par les poids $\hat{\tau}_i^{-1}/E_s(\hat{\tau}_i^{-1})$ de manière à tenir compte des effets nets de l'échantillonnage sur les fdp . Conditionnelles $f(y \mid x)$, et de la même manière d'utiliser les poids q dans $\hat{\beta}_q$. Remarquons que tandis que les poids d'échantillonnage w_i peuvent dépendre de y , x et possiblement de d'autres variables, les poids $\hat{\tau}_i^{-1}$ dépendent seulement de y et x . La mise en application de cette idée n'affecte pas le biais mais les E.-T. empiriques des estimateurs modifiés sont 0,151 et 0,053, plus petits que les E.-T. de $\hat{\beta}_{\text{sel}}$ mais semblables à ceux de $\hat{\beta}_q$.

Si l'on examine les propriétés des estimateurs de variance, le premier résultat remarquable est que les estimateurs de variance sous randomisation et sous le modèle d'échantillon (équations 4.4 et 4.5) sont fort semblables pour chaque estimateur des coefficients de régression, même s'ils sont calculés fort différemment. Pour $\hat{\beta}_{\text{ols}}$, $\hat{\beta}_{\text{pw}}$ et $\hat{\beta}_q$, les estimateurs de variance sont presque sans biais, mais pour les autres estimateurs, les estimateurs de variance sous-estiment la variance réelle. Cela tient au fait que ces estimateurs de variance ne tiennent pas compte de certaines opérations intervenant dans le calcul des coefficients de régression estimés. Donc, dans le cas des estimateurs $\hat{\beta}_{\text{mg}}$ et $\hat{\beta}_{\text{mg-q}}$, les estimateurs de variance ne tiennent pas compte

Enfin, les estimateurs de la variance par le bootstrap pour processus indépendamment B fois, l'estimateur de la variance par le bootstrap est

$$\text{Var}_{BS}(\hat{\beta}) = \frac{1}{B} \sum_{b=1}^B (\hat{\beta}^{(b)} - \bar{\hat{\beta}})(\hat{\beta}^{(b)} - \bar{\hat{\beta}})' ; \quad \bar{\hat{\beta}} = \frac{1}{B} \sum_{b=1}^B \hat{\beta}^{(b)} \quad (4.6)$$

où $\hat{\beta}$ représente n'importe lequel des estimateurs définis par 4.1.1 à 4.1.8 et $\hat{\beta}^{(b)}$ est l'estimateur correspondant calculé pour l'échantillon bootstrap b , $b = 1, \dots, B$.

4.3 Calcul des intervalles de confiance

Nous considérons deux approches du calcul des intervalles de confiance (IC) de niveau $(1 - \alpha)$. La première approche est celle de l'IC classique,

$$\hat{\beta}_k \pm Z_{1-\frac{\alpha}{2}} \hat{s.e.}(\hat{\beta}_k), \quad k = 0, 1,$$

où $\hat{\beta}_k$ représente n'importe lequel des estimateurs considérés et $\hat{s.e.}(\hat{\beta}_k)$ est l'estimateur correspondant de l'erreur-type obtenu par l'une de méthodes énumérées plus haut. La deuxième approche, du « bootstrap ordinaire », consiste à utiliser les quantiles $bs(k, \alpha)$ des estimateurs bootstrap $\hat{\beta}_k^{(b)}$ pour calculer l'IC

$$\left[2\hat{\beta}_k - bs\left(k, 1 - \frac{\alpha}{2}\right), 2\hat{\beta}_k - bs\left(k, \frac{\alpha}{2}\right) \right], \quad k = 1, 2.$$

Nous avons également essayé d'utiliser la « méthode du bootstrap studentisé », mais les taux de couverture n'étaient meilleurs pour aucun des estimateurs $\hat{\beta}_k$. Voir la remarque 14 plus loin.

4.4 Résultats des simulations

Le tableau 1 donne les moyennes empiriques des estimations énumérées à la section 4.1 sur les 2 000 populations et échantillons, et les erreurs-types (E.-T.) empiriques correspondantes. Sont également présentées les racines carrées des moyennes des estimations de variance obtenues en estimant la variance sous randomisation (« Ran. ») et en estimant la variance sous le modèle d'échantillon (« M.E. »). En raison de contraintes de temps de calcul, les résultats des estimateurs bootstrap de la variance (« BS ») sont fondés sur 300 échantillons bootstrap tirés de chacun des 500 échantillons originaux. Ces nombres d'échantillons originaux et d'échantillons bootstrap ont produit des estimateurs de variance stables.

Un estimateur de variance sous le modèle de l'échantillon qui tient compte de l'hétéroscédasticité possible est

donné par

$$\text{Var}_{\text{sm}}(\hat{\beta}_i) = [X_s' W_s T_s X_s]^{-1} \left[\sum_{i \in s} w_i^2 t_i^2 \hat{e}_{hi}^2 x_i x_i' \right] [X_s' W_s T_s X_s]^{-1}, \quad (4.5)$$

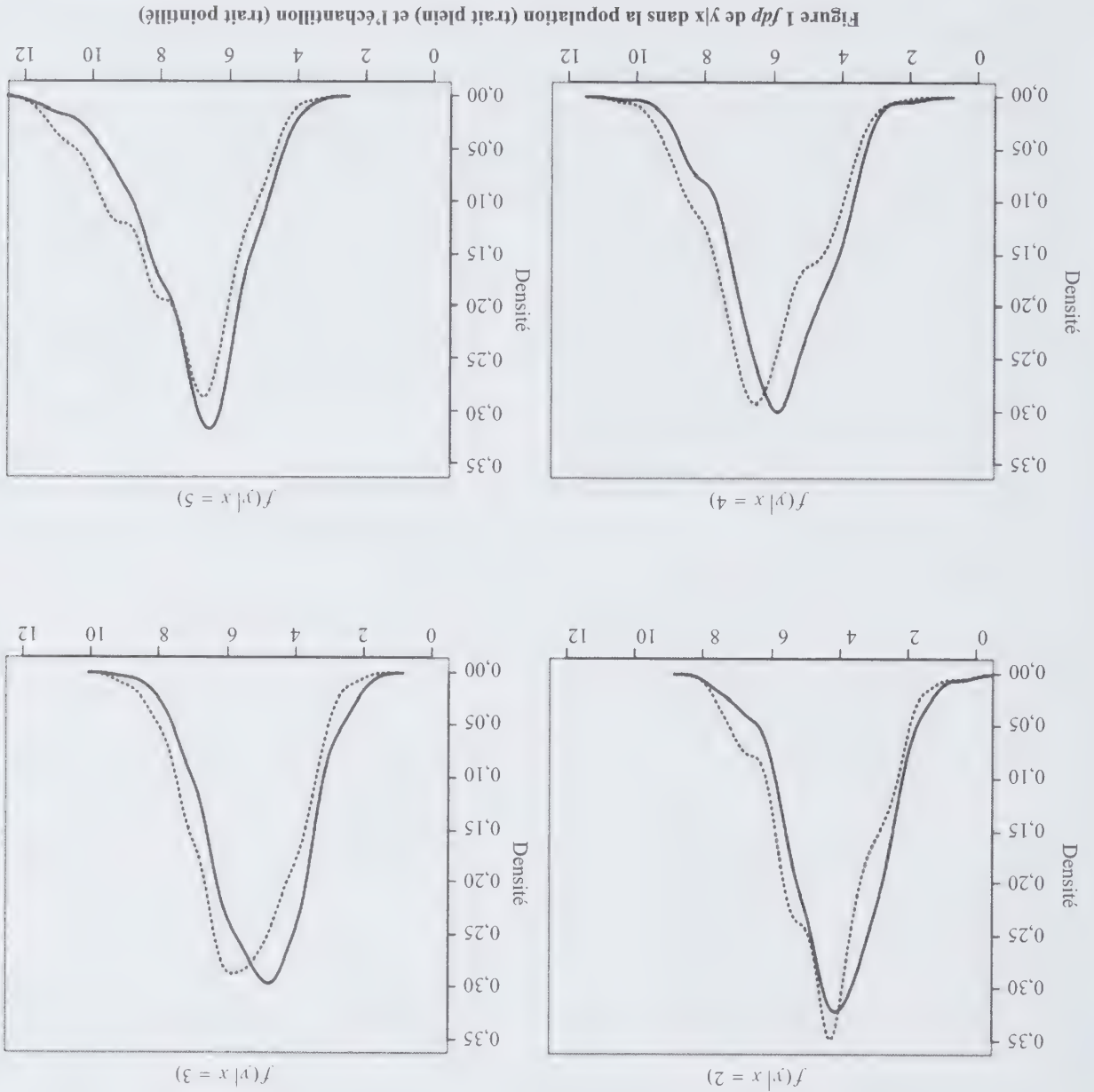
où $\hat{e}_{hi} = (y_i - x_i' \hat{\beta}_i)$. Les estimateurs de variance sous randomisation et sous le modèle d'échantillon pour l'estimateur d'échantillon avec la fonction de vraisemblance définie par (4.1), nous estimons seulement la variance sous le modèle d'échantillon en nous servant de l'inverse de la matrice d'information.

en supposant que l'échantillonnage est fait avec remise dans la strate.

$$\begin{aligned} \text{Var} \left[\sum_{i=1}^n w_i t_i x_i' e_{hi} \right] &= \sum_{h=1}^5 \text{Var} \left(\sum_{j=1}^{n_h} w_{hj} \tilde{e}_{hj,i} \right) \\ &= \sum_{h=1}^5 \frac{(n_h - 1)}{n_h} \sum_{j=1}^{n_h} (w_{hj} \tilde{e}_{hj,i} - \bar{e}_{hi})(w_{hj} \tilde{e}_{hj,i} - \bar{e}_{hi})', \end{aligned} \quad (4.4)$$

où $\tilde{e}_{hj,i} = t_{hj} x_{hj} (y_{hj} - x_{hj}' \hat{\beta}_i)$ et

$$\bar{e}_{hi} = \frac{1}{n_h} \sum_{j=1}^{n_h} w_{hj} \tilde{e}_{hj,i}$$



α dans l'intervalle $[-2, 2]$ qui minimise le déterminant de l'estimateur de variance asymptotique (3.12).

4.1.5 L'estimateur β^q est défini par (3.16). Dans la présente étude, nous n'émettons l'hypothèse d'aucun modèle paramétrique pour l'espérance $E_s(w_i | x_i)$ dans le dénominateur de q_i et estimons $\hat{E}_s(w_i | x_i) = \bar{w}_s(x_i)$, la moyenne des poids d'échantillonnage observés pour les unités avec $x = x_i$.

4.1.6 L'estimateur q -pondéré modifié β^{mg-q} est défini par (3.17). Les poids q_i sont obtenus comme dans 4.1.5 et les fonctions $a_i^q(\alpha)$, comme dans 4.1.4.

4.1.7 Les estimateurs dérivés par maximisation de la fonction de vraisemblance dans l'échantillon (3.19). L'utilisation de cette approche requiert que l'on spécifie la *fdp* de la population et l'espérance $E_s(w_i | y_i, x_i)$. Les paramètres inconnus du modèle de la population sont $\theta' = (\beta, \sigma^2)$ et nous supposons que $f^p(y_i | x_i; \theta) = N(x_i' \beta, \sigma^2)$, ce qui, comme il est noté plus haut et illustré à la figure 1, n'est pas la *fdp* correcte puisque les coefficients aléatoires ζ_j ne sont pas normaux (voir la section 3.1). Nous avons estimé $E_s(w_i | y_i, x_i; \gamma)$ non paramétriquement et établi la fonction de vraisemblance comme il suit :

Soit s_{x_i} l'échantillon des unités avec $x = x_i$ de taille m_{x_i} . Nous commençons par diviser l'échantillon en $c(x_i)$ grappes homogènes en nous fondant sur les valeurs croissantes de la variable résultat y en utilisant la fonction « hlist » du logiciel R. Les valeurs de $c(x_i)$ sont comprises entre 1 et 7, selon la taille m_{x_i} de l'échantillon (une grappe si $m_{x_i} \leq 10$, 2 grappes si $m_{x_i} \leq 20$, ..., 7 grappes si $m_{x_i} \geq 70$). Soit $b_{x_i,k}$ le point médian entre la valeur de y la plus élevée dans la grappe k et la valeur la plus faible de y dans la grappe $(k+1)$, $k = 1, \dots, c(x_i) - 1$, et définissons $b_{x_i,0} = -\infty$, $b_{x_i,c(x_i)} = +\infty$. Pour $b_{x_i,k-1} \leq y \leq b_{x_i,k}$, nous avons estimé $E_s(w_i | y_i, x_i)$ par la moyenne $\bar{w}_s(y_i, x_i) = \bar{w}_s(x_i)$ des poids d'échantillonnage des unités dont les valeurs de y sont comprises dans le même intervalle. La substitution de $E_s(w_i | y_i, x_i) = \bar{w}_s(y_i, x_i)$ dans (3.19) définit la fonction de vraisemblance d'échantillon utilisée pour la présente étude en simulation comme étant

$$L_s(\theta; y_s, x_s) = \prod_{i \in s} \frac{f^p(y_i | x_i; \theta) / \bar{w}_s(y_i, x_i)}{[F^p(b_{x_i,k(x_i)}) - F^p(b_{x_i,k(x_i)-1})] / \bar{w}_s(x_i)}, \quad (4.1)$$

où $F^p(b_{x_i,k(x_i)}) = \int_{-\infty}^{b_{x_i,k(x_i)}} f^p(y | x_i; \theta) dy$ (la fonction de cumulative de la *fdp* supposée normale). L'approximation (4.1) est semblable à l'approximation (3.20) proposée pour le cas où x ainsi que y sont des variables discrètes.

Remarque 13. Afin de faciliter les optimisations numériques utilisées pour le calcul des estimateurs β^{mg} , β^{mg-q} et les

4.2 Estimation de la variance

Nous avons appliqué trois approches d'estimation de la variance. La première consiste à estimer la variance sous randomisation, la deuxième, à estimer la variance sous le modèle d'échantillon non paramétrique qui, de la même façon estime la variance sous le modèle d'échantillon. Considérons d'abord les estimateurs définis par 4.1.1, 4.1.3 à 4.1.6 et 4.1.8 à la section 4.1. Tous ces estimateurs peuvent s'écrire sous la forme générique.

$$\hat{\beta}_i' = \left[\sum_{i=1}^n w_i^t x_i x_i' \right]^{-1} \sum_{i=1}^n w_i^t x_i y_i \quad (4.2)$$

où $X_i' = [x_1, \dots, x_n]$, $W_s = \text{diag}[w_1, \dots, w_n]$ est la matrice diagonale ayant le poids d'échantillonnage sur la diagonale principale et $T_s = \text{diag}[t_1, \dots, t_n]$ avec les t_i définis par les estimateurs. Pour β_{ols} , $t_i = 1 / w_i$, pour β_{sel} , $t_i = w_i^{-1-1/t_i}$ et ainsi de suite. La variance sous randomisation de ces estimateurs et estimée par

$$\text{Var}(\hat{\beta}_i') = [X_i' W_s T_s X_i']^{-1} \left[\text{Var} \sum_{i=1}^n w_i^t x_i e_i'' \right] [X_i' W_s T_s X_i']^{-1}, \quad (4.3)$$

où $e_i'' = (y_i - x_i' \beta)$ et B est l'estimateur sous recensement. En utilisant le double indice (hi) pour définir la j^{e} unité dans l'échantillon s_h de taille n_h tiré de la strate h , nous estimons

Nous avons généré les valeurs de population d'une covariable discrète unique x en commençant par générer les observations \tilde{x}_j au moyen d'une loi *gamma* de moyenne 2 et de variance 4, puis nous avons défini x_j comme étant l'entier le plus proche de \tilde{x}_j si $\tilde{x}_j < 5$ et $x_j = 5$ autrement. Les covariables sont par conséquent $x_j = (1, x_j)'$, avec $x_j = 0, 1, \dots, 5$. Les covariables de population ont été générées une fois et maintenues fixes pour toutes les populations. La figure 1 montre les *fdp* de la population et de l'échantillon de la variable résultat y pour $x = 2, 3, 4, 5$. On peut voir que les *fdp* de la population et de l'échantillon diffèrent, ce qui indique que le processus d'échantillonnage est informatif. Notons aussi que la *fdp* de la population n'est pas normale parce que les coefficients aléatoires ζ_j ne sont pas normaux. Nous étudions la performance des diverses méthodes en examinant le biais, la variance, l'estimation de la variance et la couverture des intervalles de confiance. Nous supposons pour toutes les méthodes que les seules informations disponibles sont les résultats observés et les covariables (y^{hs}, x^{hs}) pour chaque strate h , les probabilités de sélection dans l'échantillon et les tailles réelles de strates $\{N_h\}$. Selon nous, cela correspond à la pratique dans la plupart des applications réelles.

4.1 Estimateurs examinés

4.1.1 L'estimateur par les moindres carrés ordinaires $\hat{\beta}_{ols}$. L'utilisation de cet estimateur ne tient pas compte du processus d'échantillonnage.

4.1.2 L'estimateur proposé par Feder (2011, voir la section 3.2). L'application de cette approche se fait en quatre

étapes : i) ajuster un modèle linéaire avec une variance résiduelle constante dans chaque strate, ii) imputer les valeurs manquantes des covariables pour les unités non échantillonnées en échantillonnant avec remise $(N_h - n_h)$ valeurs parmi les n_h valeurs observées dans la strate h avec les probabilités $\tilde{p}_{hi} = (w_{hi} - 1) / \sum_{i=1}^{n_h} (w_{hi} - 1)$ à chaque tirage, où les w_{hi} sont les poids d'échantillonnage quand on échantillonne la strate h , iii) imputer les valeurs manquantes de y dans chaque strate en générant des observations au hasard à partir du modèle ajusté à l'étape i), iv) ajuster le modèle de régression linéaire de y sur x en utilisant les données de population avec les valeurs manquantes pour les unités non échantillonnées remplacées par les valeurs imputées. Nous désignons l'estimateur résultant par $\hat{\beta}_f$.

4.1.3 L'estimateur PW (pondéré par les probabilités) $\hat{\beta}_{pw}$ (équation 3.8).

4.1.4 L'estimateur $\hat{\beta}_{mg}$ proposé par Magee (1998, voir la section 3.5). Dans notre application, nous définissons $a_i(\alpha) = (x_i + 0, 1)'$ et recherchons la puissance optimale

L'estimateur $\hat{\beta}_{sel}$ a la même forme que l'estimateur PW $\hat{\beta}_{pw}$ dans (3.8), mais avec les poids $\tau_i^{-1} = 1 / \Pr(i \in s | y_i, x_i)$ au lieu des poids d'échantillonnage w_i . En pratique, on doit remplacer les probabilités τ_i par les estimations sur l'échantillon $\hat{\tau}_i$. Voir la section 4.

calage de la forme

$$\sum_{i=1}^n p_i (\pi_i - \hat{\tau}_i) k(y_i, x_i) = 0 \quad (3.34)$$

pour améliorer l'estimation des probabilités $\{p_i\}$ dans (3.31), où $k(y_j, x_j) = k(g_j)$ est une fonction du résultat observé et des covariables. Des exemples de fonctions plausibles pour le cas d'une covariable unique x sont $k(g_j) = y_j x_j$, $k(g_j) = y_j / x_j$, etc. La caractéristique importante des contraintes (3.34) est qu'elle ne nécessite pas la connaissance des quantités de population telles que les moyennes des variables de calage, comme il est souvent supposé lorsque l'on recommande l'approche de la vraisemblance empirique pour l'estimation d'après des données d'enquête. Clairement, quand les moyennes \bar{C}_U des variables de calage sont connues, des contraintes de la forme $\sum_{i=1}^n p_i c_i = \bar{C}_U$ peuvent être ajoutées également. (Voir aussi la remarque 14).

4. Étude empirique

Dans la présente section, nous présentons les résultats d'une étude en simulation destinée à évaluer et à comparer les propriétés des méthodes discutées à la section 3. Les conditions de simulation sont décrites à la section 3.1 et nous utilisons $H = 5$ strates. Les paramètres cibles sont les coefficients de régression $\beta' = (\beta_0, \beta_1) = (2, 1)$ de l'espace de population (3.1). L'étude en simulation consiste à générer 2 000 populations et échantillons (un échantillon pour chaque population) et à calculer les estimateurs, les estimateurs de variance et les intervalles de confiance énumérés plus bas pour chaque échantillon. La taille de la population est de 5 000 avec des tailles approximatives de strates de $N_h = 363, 554, 842, 1 278, 1 963$. (Les tailles de strates sont aléatoires.) La taille d'échantillon est $n = 300$ avec $n_h = 60$ unités échantillonnées dans chaque strate. Les fractions d'échantillonnage varient par conséquent fortement d'une strate à l'autre.

n'est pas nécessairement simple non plus selon le modèle de population.

Remarque 11. L'utilisation de la méthode du principe de l'information manquante dans les conditions de simulation de la section (3.1) requiert que l'on connaisse les covariables et l'appartenance aux strates des unités non comprises dans l'échantillon. Nous n'avons pas trouvé de moyen d'appliquer la méthode dans ce cas sans formuler des hypothèses supplémentaires concernant la distribution conjointe des covariables et des variables du plan d'échantillonnage.

3.6.3 Fonction de vraisemblance empirique

Ces dernières années, on a assisté à un gain d'intérêt pour

les méthodes fondées sur la fonction de vraisemblance empirique (EL, pour *empirical likelihood*) pour analyser les données d'enquêtes complexes. La méthode EL proposée originellement par Hartley et Rao (1968) dans le contexte des sondages et par Owen (1988, 2001) combine la robustesse des méthodes non paramétriques avec l'efficacité de l'approche fondée sur la fonction de vraisemblance. Deux autres avantages importants de cette méthode tiennent au fait qu'elle se prête très naturellement à l'utilisation des équations de calage et qu'elle permet de construire des intervalles de confiance sans qu'il soit nécessaire d'estimer la variance.

Considérons le modèle défini par (3.13) où, pour le moment, nous considérons les covariables comme étant aléatoires, et notons $g_i = (y_i, x_i)'$. Sous certaines conditions de régularité, le paramètre vectoriel θ est la solution unique de l'équation

$$E_p \left\{ \frac{\partial \log L(\theta)}{\partial \theta} [y - m(x; \theta)] \right\} = 0.$$

Soit p_1, \dots, p_n un ensemble de probabilités correspondant aux observations (g_1, \dots, g_n) , tel que p_i est le « saut » (masse de probabilité) de la fonction de répartition de la population $F_p(g_i)$ à g_i . L'hypothèse est que F_p s'appuie sur les valeurs observées de sorte que

$$\sum_{i=1}^n p_i \frac{\partial \log L(x_i; \theta)}{\partial \theta} [y_i - m(x_i; \theta)] = 0. \quad (3.29)$$

En supposant que les observations sont indépendantes, la vraisemblance empirique de F_p est $L(F_p) = \prod_{i=1}^n p_i$. Notons que, si p_i est une fonction connue de certains paramètres inconnus, $L(F_p)$ coïncide avec la fonction de vraisemblance paramétrique classique. Les estimateurs EL (non paramétrique) des probabilités p_i sont la solution $p_i^{(p)}$ du problème de maximisation

$$\max_{p_1, \dots, p_n} \prod_{i=1}^n p_i \quad \text{s.c.} \quad p_i \geq 0, \sum_{i=1}^n p_i = 1, \quad (3.30)$$

donnant $p_i^{(p)} = 1/n, i = 1, \dots, n$. Dans le cas de la régression linéaire, $m(x_i; \theta) = x_i' \beta$ et, en substituant $p_i^{(p)}$ à p_i dans (3.29) et en résolvant les équations, nous obtenons l'estimateur EL de β comme étant $\hat{\beta}_{\text{OLS}} = \hat{\beta}_{\text{OLS}}$. Si les moyennes de population finie \bar{C} des variables C mesurées dans l'échantillon sont connues, elles peuvent être ajoutées au problème de maximisation (3.30) en ajoutant les contraintes de calage $\sum_{i=1}^n p_i c_i = \bar{C}$. Cette information supplémentaire devrait améliorer l'estimation des p_i et donc l'estimation des paramètres inconnus du modèle. Voir aussi la remarque 12 qui suit.

Supposons maintenant que les unités sont tirées de l'échantillon (ou qu'elles répondent) avec des probabilités de sélection inégales π_i . Dans ce cas, il est fréquent de remplacer la fonction de vraisemblance empirique objectif $L(F_p) = \prod_{i=1}^n p_i$ par la fonction de vraisemblance pseudo-empirique $L_{\text{pi}}(F_p) = \prod_{i=1}^n p_i w_i$, où, comme auparavant, $w_i = 1/\pi_i$. Notons que $\log L_{\text{pi}}(F_p) = \sum_{i=1}^n w_i \log(p_i)$ est l'estimateur HT de $\log L^{\text{pop}}(F^p) = \sum_{i=1}^N \log p_i$. Les estimateurs de la pseudo-vraisemblance empirique des p_i résolvent le problème de maximisation,

$$\max_{p_1, \dots, p_n} \prod_{i=1}^n p_i w_i \quad \text{s.c.} \quad p_i \geq 0, \sum_{i=1}^n p_i = 1. \quad (3.31)$$

Voilà, par exemple, Chen et Sitter (1999). Il est facile de vérifier qu'en l'absence de contraintes d'échantillonnage, la solution de (3.31) est $p_i^{(\text{pel})} = w_i / \sum_{i=1}^n w_i$, et en substituant $p_i^{(\text{pel})}$ à p_i dans (3.29), que $\hat{\beta}_{\text{pel}} = \hat{\beta}_{\text{pw}}$, l'estimateur PW (3.8).

Les fonctions de vraisemblance empirique (3.30) et (3.31) sont établies par rapport à la distribution de la vraie population. On peut aussi tenir l'estimateur de la vraisemblance empirique en définissant la vraisemblance par rapport à la distribution de l'échantillon $f_s(g_i) = \Pr(I_i = 1 | g_i) f_p(g_i) / \Pr(I_i = 1)$, où, en notant $\tau_i = \Pr(I_i = 1 | g_i)$, $\Pr(I_i = 1) = \sum_{i=1}^n p_i \tau_i$. En s'inspirant de Kim (2009) et de Chaudhuri, Handcock et Rendall (2010), on obtient les estimateurs de vraisemblance empirique EL des probabilités p_i comme la solution du problème de maximisation

$$\max_{p_1, \dots, p_n} \left[\sum_{i=1}^n \log(p_i \tau_i) - n \log \sum_{i=1}^n p_i \tau_i \right]$$

$$\text{s.c.} \quad p_i \geq 0, \sum_{i=1}^n p_i = 1. \quad (3.32)$$

La solution de (3.32) est $p_i^{\text{sel}} = \tau_i^{-1} / \sum_{j=1}^n \tau_j^{-1}$ et en l'introduisant par substitution dans (3.29),

plus bas) requiert la connaissance des covariables des unités non échantillonnées.

3. L'application de cette approche permet d'utiliser l'inférence conditionnelle, sachant l'échantillon d'unités répondantes, par exemple, en conditionnant sur les covariables observées.

4. Les modèles vérifiés pour les résultats observés et les probabilités de réponse définissent le modèle à vérifier pour les résultats manquants des unités non échantillonnées ou des non-répondants, qui peut être utilisé pour l'imputation de ces résultats. Les méthodes fondées sur la pondération probabiliste et leurs variantes permettent d'estimer le modèle de la non-réponse NMAR, le modèle de la population ne peut être utilisé pour la prédiction ni pour l'imputation des résultats manquants. Voir Sverchkov et Pfeffermann (2004) et Pfeffermann et Sikov (2011) pour des illustrations.

5. L'utilisation du modèle d'échantillon permet de vérifier si le processus d'échantillonnage peut être ignoré. Pfeffermann et Sverchkov (2009) passent la littérature pour vérifier si l'on peut ignorer la sélection de l'échantillon.

3.6.2 La fonction de vraisemblance complète

Théoriquement, un moyen plus efficace d'estimer les paramètres inconnus du modèle de la population consiste à fonder la fonction de vraisemblance sur la distribution conjointe des données d'échantillon et des indicateurs d'appartenance à l'échantillon. Sous réponse complète, la fonction de *vraisemblance complète* est alors

$$L_f(\theta, \gamma; I_U, y_s, x_s, x_s) = \prod_{i \in s} \Pr(I_i = 1 | y_i, x_i; \gamma) f^p(y_i | x_i; \theta)$$

$$\prod_{j \in s} [1 - \Pr(I_j = 1 | x_j; \theta, \gamma)] \quad (3.27)$$

où $I_U = \{I_1, \dots, I_N\}$ est le vecteur des indicateurs d'inclusion dans l'échantillon et $\Pr(I_j = 1 | x_j; \theta, \gamma) = \int \Pr(I_j = 1 | y_j, x_j; \gamma) f^p(y_j | x_j; \theta) dy_j$ est le *score de propension* à répondre de l'unité j . La fonction de vraisemblance (3.27) suppose que $\Pr(I_{-U} | y_U, x_U) = \prod_{k \in U} \Pr(I_k | y_k, x_k)$ (échantillonnage de Poisson), mais elle peut être généralisée à d'autres plans d'échantillonnage. La fonction de vraisemblance complète à l'avantage de tenir compte des probabilités d'échantillonnage des unités non comprises dans l'échantillon, donc d'utiliser plus d'information, mais elle requiert que l'on connaisse les covariables de toutes les unités de la population. Voir, par exemple, Gelman, Carlin, Stern et Rubin (2003) et Little (2004). La modélisation de la

distribution conjointe des covariables pour les unités non comprises dans l'échantillon et l'élimination de ces covariables de la fonction de vraisemblance par intégration peut être très compliquée en pratique et constituer un vrai tour de force quand elles sont nombreuses. Pfeffermann (2006) comparent empiriquement l'utilisation de la fonction de vraisemblance de l'échantillon avec celle de la fonction de vraisemblance complète pour des modèles multiniveaux dans un contexte bayésien. Les deux approches produisent des résultats comparables, mais naturellement, cela pourrait ne pas être le cas dans d'autres applications.

Un autre moyen de définir la fonction de vraisemblance complète consiste à appliquer le principe de l'*information manquante* (Orchard et Woodbury 1972). L'idée fondamentale est d'exprimer la fonction de score dans l'échantillon comme étant l'espérance conditionnelle de la fonction de score dans la population, sachant les données d'échantillon. À l'instar de Chambers et Skinner (2003, chapitre 2), définissons la fonction de *vraisemblance complète de l'échantillon* par $L_{f^p}(\lambda) = f(\lambda; y_s, x_s, I_U, z_U)$, où comme auparavant, z_U est une matrice connue des valeurs de population qui sous-tend la sélection de l'échantillon et λ définit les paramètres inconnus du modèle. La fonction de *vraisemblance complète de la population* correspondante est $L_{f^p}(\lambda) = f(\lambda; y_U, x_U, I_U, z_U)$, où $y_U = (y^s, y^g)$ et $x_U = (x^s, x^g)$. Le principe de l'information manquante énonce que

$$sc_g(\lambda) = (\partial / \partial \lambda) \log L_{f^p}(\lambda)$$

$$= E_p[(\partial / \partial \lambda) \log L_{f^p}(\lambda) | y_s, x_s, I_U, z_U]. \quad (3.28)$$

Une autre identité définit la relation entre la matrice d'information pour la fonction de population et la matrice d'information pour la fonction de vraisemblance de l'échantillon.

Breckling, Chambers, Dorfman, Tam et Welsh (1994) et Chambers et coll. (1998) considèrent les applications du principe de l'information manquante aux données d'enquêtes complexes. En particulier, Chambers et coll. (1998) étudient l'utilisation de ce principe lorsqu'on ne dispose que de renseignements limités sur le plan d'échantillonnage au lieu de l'information complète comprise dans z_U . Les auteurs donnent des exemples où l'utilisation du principe de l'information manquante est plus efficace que celle de la fonction de vraisemblance de l'échantillon $L_s(\theta, \gamma; y_s, x_s)$ défini par (3.19), qui n'emploie que les poids $\{w_i, i \in s\}$. La vraisemblance (3.28) peut être étendue afin de tenir compte de la non-réponse NMAR, mais l'application de cette approche requiert que l'on connaisse alors les valeurs de population des variables qui expliquent la réponse. Le calcul de l'espérance dans le deuxième membre de (3.29)

Jusqu'à présent, nous avons supposé que la réponse était complète. Considérons maintenant le cas d'une non-réponse de type NMAR. Dans ces conditions, le processus de réponse doit être modélisé également. En vertu de (2.2) et avec une notation supplémentaire des paramètres, la fonction de vraisemblance des « répondants » prend la forme

$$L_o = \prod_{i=1}^I f(y_i | x_i, I_i = 1, R_i = 1; \theta^*, \gamma^*)$$

$$= \prod_{i=1}^I \frac{\Pr(R_i = 1 | y_i, x_i, I_i = 1; \gamma^*) f_s(y_i | x_i; \theta^*)}{\Pr(R_i = 1 | x_i, I_i = 1; \gamma^*, \theta^*)}, \quad (3.25)$$

où $\theta^* = (\theta, \gamma)$ représente les paramètres de la distribution dans l'échantillon sous réponse complète (équation 3.19), et γ^* représente les paramètres du processus de réponse. Notons que, contrairement aux probabilités d'échantillonnage $\pi_i = \Pr(i \in s)$, qui sont généralement connues et peuvent être utilisées pour estimer les probabilités $\Pr(I_i = 1 | y_i, x_i, \gamma)$ comme nous l'avons expliqué plus haut, les probabilités de réponse sont généralement inconnues.

Chang et Kott (2008) proposent une méthode d'estimation des probabilités de réponse dans laquelle ils utilisent les totaux connus des variables de calage. Les auteurs émettent l'hypothèse d'un modèle paramétrique pour les probabilités de réponse qui peut dépendre de la valeur observée du résultat, et estiment les paramètres inconnus de ce modèle en effectuant la régression des totaux des variables de calage en fonction des estimateurs HT. Les poids utilisés pour les estimateurs HT sont égaux au produit des poids d'échantillonnage et de l'inverse des probabilités de réponse sous le modèle. Soit c_i définit les valeurs des variables de calage pour l'unité i et denote $p(y_i, x_i; \gamma^*) = \Pr(R_i = 1 | y_i, x_i, I_i = 1; \gamma^*)$. Chang et Kott (2008) estiment les paramètres inconnus en établissant les équations de régression non linéaires

$$C_U = \sum_i w_i \frac{p(y_i, x_i; \gamma^*)}{c_i} + \varepsilon^*,$$

où $C_U = \sum_{i=1}^N c_i^*$ et ε^* est un vecteur des erreurs. Les paramètres γ^* sont estimés au moyen de l'algorithme itératif

$$\hat{\gamma}^{(j+1)} = \hat{\gamma}^{(j)} + \left\{ \hat{H}(\hat{\gamma}^{(j)})^T V^{-1}(\hat{\gamma}^{(j)}) \hat{H}(\hat{\gamma}^{(j)}) \right\}^{-1}$$

$$\left(C_U - \sum_i w_i \frac{\pi(y_i, v_i; \hat{\gamma}^{(j)})}{c_i} \right) \quad (3.26)$$

où

calculé pour $\gamma = \hat{\gamma}^{(j)}$.

Chang et Kott (2008) n'émettent pas l'hypothèse d'un modèle pour le résultat et leur approche est par conséquent limitée à l'estimation du modèle pour les probabilités de réponse. Pfeffermann et Sikov (2011) utilisent la fonction de vraisemblance (3.25) pour estimer les modèles de la population en supposant que l'échantillonnage n'est pas informatif. La maximisation de la fonction de vraisemblance est effectuée par itération entre la maximisation de la vraisemblance par rapport à θ^* quand γ^* est donné et la solution des équations de calage par rapport à γ^* quand θ^* est donné. Les auteurs montrent comment estimer les covariables et la variable résultat manquante pour une unité non-répondante et utiliser cette distribution pour imputer les résultats manquants, donc prédire le total de population finie de la variable résultat.

L'estimation du modèle de population en ajustant le modèle d'échantillon présente des avantages que n'offrent pas les autres approches considérées dans le présent article.

1.

Une fois que le modèle d'échantillon est spécifié, il permet de procéder à une inférence classique fondée sur un modèle tel que les méthodes fondées sur la fonction de vraisemblance, l'inférence bayésienne ou la modélisation semi-paramétrique. Il est important d'insister à cet égard sur le fait que l'ajustement du modèle de population peut être évalué en testant l'adéquation de l'ajustement du modèle d'échantillon ajusté sur les résultats observés, en utilisant les techniques classiques de diagnostic de modèle. Voir Krieger et Pfeffermann (1997) et Pfeffermann et Sikov (2011) pour des statistiques de test appropriées et des exemples.

2. La fonction de vraisemblance de l'échantillon donne un moyen cohérent de traiter la non-réponse NMAR quand on estime des modèles de population. Les méthodes fondées sur la pondération probabiliste requièrent la connaissance des probabilités de réponse ou de bons estimateurs de celles-ci. L'utilisation de la fonction de vraisemblance complète (voir

Estimer le paramètre vectoriel γ en dehors de la fonction de vraisemblance, puis introduire l'estimation dans (3.19) et maximiser la vraisemblance comme une fonction du paramètre vectoriel θ seulement produit habituellement des résultats plus stables que maximiser la vraisemblance sur (θ, γ) simultanément.

L'estimation des espérances $E_s(w_i | \gamma_i, x_i; \gamma)$ et $E_s(w_i | x_i; \theta, \gamma)$ dans le cas de probabilités d'inclusion discrètes est similaire.

Exemple 9. Considérons le cas d'une régression logistique multinomiale avec une covariable discrète x et M valeurs possibles du résultat y . Si l'on suppose que $E_s(w_i | y_i = m, x_i = k)$ n'est pas une fonction des paramètres du modèle, elle peut être estimée par \bar{w}_{mk} , la moyenne des poids dans la cellule (m, k) , d'où $\hat{\pi}_{mk} = \Pr_p(i \in s | y_i = m, x_i = k) = (1 / \bar{w}_{mk})$. Nous obtenons :

$$\Pr_s(y_i = m | x_i = k; \theta) \equiv \frac{[\Pr_p(y_i = m | x_i = k; \theta) / \bar{w}_{mk}]}{\sum_{i=1}^M [\Pr_p(y_i = m^* | x_i = k; \theta) / \bar{w}_{mk}]} \quad (3.20)$$

Les poids d'échantillonnage figurent dans le modèle de l'échantillon, mais il ne s'agit pas d'une application de la pondération probabiliste classique. Notons qu'avec cette approximation, les paramètres du modèle de population et ceux du modèle d'échantillon sont les mêmes. Dans notre étude empirique, nous utilisons une approximation similaire pour la distribution dans l'échantillon en catégorisant les valeurs d'une variable résultat continue. Voir Pfeffermann et Sverchokov (1999) pour d'autres exemples.

Ensuite, considérons l'estimation du paramètre vectoriel θ régissant le modèle de population. Sous des conditions faibles, θ est la solution unique des équations

$$W^U(\theta) = \sum_{j \in U} E_p(\delta_j | x_j) = 0; \quad \delta_j = (\delta_{j,0}, \delta_{j,1}, \dots, \delta_{j,K})' = \partial \log f_p(\gamma_j | x_j; \theta) / \partial \theta. \quad (3.21)$$

Pfeffermann et Sverchokov (2003) considèrent trois approches distinctes pour estimer θ . La caractéristique commune de ces approches est que les seules données utilisées pour l'estimation sont les observations $\{(y_i, x_i, w_i), i \in s\}$, comme dans le cas des estimateurs PW et de leurs modifications considérées à la section 3.5. À la section 3.6.2, nous examinons l'utilisation de la fonction de « vraisemblance complète », qui suppose que l'on connaît les covariables $\{x_j, j \in U\}$ et, éventuellement, des renseignements supplémentaires sur le plan d'échantillonnage également.

La première approche consiste à redéfinir les équations des paramètres par rapport au modèle de l'échantillon. Si l'on suppose dans (3.19) que $E_s(w_i | x_i; \theta, \gamma)$ est différentiable par rapport à θ , les équations des paramètres du

modèle d'échantillon sont $W_{1s}^s(\theta) = \sum_{i \in s} E_s\{\partial \log f_s(\gamma_i | x_i; \theta, \gamma) / \partial \theta\} = \sum_{i \in s} E_s\{\delta_i + \partial \log E_s(w_i | x_i; \theta, \gamma) / \partial \theta\} | x_i\} = 0$. Le vecteur θ est estimé sous cette approche en résolvant les équations

$$W_{1s}^s(\theta) = \sum_{i \in s} [\delta_i + \partial \log E_s(w_i | x_i; \theta, \gamma) / \partial \theta] = 0. \quad (3.22)$$

La deuxième approche consiste à appliquer la relation (3.14) aux équations des paramètres (3.21). Pour un échantillon aléatoire tiré du modèle d'échantillon, les équations sont maintenant $W_{2s}^s(\theta) = \sum_{i \in s} E_s(q_i \delta_i | x_i) = 0$, où $q_i = w_i / E_s(w_i | x_i)$. Le vecteur θ est estimé sous cette approche en résolvant les équations

$$W_{2s}^s(\theta) = \sum_{i \in s} q_i \delta_i = 0. \quad (3.23)$$

La troisième approche consiste à utiliser la propriété selon laquelle, si θ est la solution de (3.21), il est alors aussi la solution des équations $W^U(\theta) = \sum_{j \in U} E_p(\delta_j) = \sum_{i \in s} E_s(w_i \delta_i) = 0$, avec les équations d'estimation $\sum_{i \in s} E_s(w_i \delta_i) = 0$, pour un échantillon aléatoire provenant du modèle de l'échantillon, les équations des paramètres sont $W_{3s}^s(\theta) = \sum_{i \in s} w_i \delta_i = 0$.

$$W_{3s}^s(\theta) = \sum_{i \in s} w_i \delta_i = 0. \quad (3.24)$$

Notons que les équations (3.24) sont les équations de la *pseudo-vraisemblance* (remarque 7).

Remarque 9. L'utilisation des poids $q_i = w_i / E_s(w_i | x_i)$ pour estimer les paramètres du modèle de la population a été justifiée à la section 3.5 en faisant référence à l'estimation par les moindres carrés. Voir la discussion présentée dans cette section concernant la différence entre l'utilisation des poids q_i et des poids w_i . Pfeffermann et Sverchokov (1999, 2003) montrent que l'estimation de θ en résolvant les équations (3.23) donne des estimateurs dont la variance sous randomisation est plus faible que l'estimation de θ en résolvant les équations (3.24). Notons que, sous l'hypothèse d'un modèle de régression linéaire appliqué à la population, la solution de (3.24) donne l'estimateur PW (3.8), et la solution de (3.23) donne l'estimateur q -pondéré (3.16).

Remarque 10. L'utilisation du modèle d'échantillon pour l'estimation des modèles de la population multivariés est examinée dans Pfeffermann, Moura et Nascimento-Silva (2006) en utilisant l'approche bayésienne. Pfeffermann et Sverchokov (2007) ajustent des modèles multivariés pour l'estimation sur petits domaines sous échantillonnage informatif des domaines et à l'intérieur des domaines en suivant l'approche fréquentiste.

Donc, pour les vecteurs θ dans l'espace des paramètres plausibles Θ ,

$$\theta = \underset{\theta \in \Theta}{\operatorname{argmin}} \frac{1}{n} \sum_{i \in S} E_p^{\theta} \{ [Y_i - m(x_i; \theta)]^2 | x_i \}$$

$$= \underset{\theta \in \Theta}{\operatorname{argmin}} \frac{1}{n} \sum_{i \in S} E_s \{ q_i [Y_i - m(x_i; \theta)]^2 | x_i \}.$$

Le vecteur θ peut donc être estimé en résolvant le problème de minimisation,

$$\hat{\theta}^q = \underset{\theta \in \Theta}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^n q_i [Y_i - m(x_i; \theta)]^2;$$

$$\hat{q}_i = w_i / E_s(w_i | x_i). \quad (3.15)$$

L'emploi de cet estimateur requiert l'estimation de $\hat{E}_s(w_i | x_i)$ mais sous des conditions de régularité faibles, \hat{E}_s est convergent pour θ même si l'espérance $E_s(w_i | x_i)$ est spécifiée incorrectement. Voir Pfeffermann et Sverchkov (2009) et la section 4.1 du présent article pour des exemples de spécification et d'estimation de $E_s(w_i | x_i)$.

Exemple 8. Sous le modèle de régression linéaire de la population avec variance constante,

$$\hat{\beta}^q = \left[\sum_{i \in S} q_i x_i x_i' \right]^{-1} \sum_{i \in S} q_i x_i y_i. \quad (3.16)$$

Comme on peut le vérifier facilement, $\hat{\beta}^q$ est convergent sous randomisation pour les coefficients de régression sous recensement $B = [\sum_{j=1}^N x_j x_j' / E_s(w_j | x_j)]^{-1} \sum_{j=1}^N x_j y_j / E_s(w_j | x_j)$, et donc convergent sous $P - r$ pour β , même quand $E_s(w_i | x_i)$ est mal spécifiée.

La différence évidente entre l'estimateur PW $\hat{\theta}^{pw}$ et l'estimateur $\hat{\theta}^q$ est que le second utilise les poids corrigés $q_i = w_i / E_s(w_i | x_i)$. Lorsque la sélection de l'échantillon dépend seulement des covariables, le processus d'échantillonnage est ignorable. Donc, pour se protéger contre l'échantillonnage informatif, il suffit uniquement de tenir compte des effets nets d'échantillonnage sur la *fdp* conditionnelle cible de $y_i | x_i$. Cela se fait en utilisant les poids q_i . En revanche, les poids d'échantillonnage w_i rendent compte des effets de l'échantillonnage sur la distribution conjointe de (y_i, x_i) . Par conséquent, ils ont tendance à être plus variables et l'estimateur $\hat{\theta}^{pw}$ possède une plus grande variance.

Une combinaison des deux dernières modifications est également possible et est examinée à la section 4. L'idée simple proposée par Moshe Feder (communication privée) est d'appliquer la modification de Magee (1998) à l'estimateur $\hat{\beta}^q$ au lieu de l'estimateur $\hat{\beta}^{pw}$, c'est-à-dire d'utiliser l'estimateur

$$L_s(\theta; \gamma; y_s, x_s) = \prod_{i \in s} \frac{E_s(w_i | x_i; \theta, \gamma) f_p(y_i | x_i; \theta)}{E_s(w_i | y_i, x_i; \gamma; \theta)}. \quad (3.19)$$

Les espérances dans le deuxième membre de (3.19) sont calculées par rapport à la *fdp* des poids d'échantillonnage dans l'échantillon. Donc, quand on connaît les poids pour les unités échantillonnées, comme cela est généralement le cas sous réponse complète, les espérances peuvent être modélisées et estimées par la régression de w_i en fonction de (y_i, x_i) , en utilisant les procédures classiques d'ajustement du modèle. Supposons pour commencer que les poids sont continus, comme dans l'échantillonnage avec probabilité proportionnelle à la taille (PPT) avec une variable de taille continue. Pour une forme donnée du modèle de population, les espérances $E_s(w_i | y_i, x_i; \gamma; \theta)$ et $E_s(w_i | x_i; \gamma; \theta)$ peuvent être obtenues en deux étapes :

1. identifier et estimer $E_s(w_i | y_i, x_i; \gamma) = E_s(w_i | y_i, x_i; \hat{\gamma})$, en utilisant les données d'échantillon.
2. intégrer $\int [1 / E_s(w_i | y_i, x_i; \hat{\gamma})] f_p(y_i | x_i; \theta) dy_i$ pour obtenir $E_p(\pi_i | x_i; \theta; \hat{\gamma})$. Calculer, $E_s(w_i | x_i; \theta; \hat{\gamma}) = 1 / E_p(\pi_i | x_i; \theta; \hat{\gamma})$ (découle de 3.14).

$$L_s(\theta; \gamma; y_s, x_s) = \prod_{i \in s} \frac{\Pr(I_i = 1 | x_i, y_i; \gamma; \theta)}{\Pr(I_i = 1 | x_i; \gamma; \theta)}. \quad (3.18)$$

Comme auparavant, nous supposons que $\Pr(I_i = 1 | \pi_i, y_i, x_i) = \pi_i$, ce qui implique que $\Pr(I_i = 1 | x_i, y_i) = E_p(\pi_i | x_i, y_i)$. En vertu de (3.14), la fonction de vraisemblance dans l'échantillon peut donc s'écrire

Un moyen naturel d'estimer les paramètres du modèle de population consiste à maximiser la fonction de vraisemblance de l'échantillon. Supposons d'abord que la réponse est complète et que les observations sur l'échantillon sont indépendantes sous la distribution dans l'échantillon. La fonction de vraisemblance prend alors la forme

3.6.1 Utilisation du modèle d'échantillon pour

l'estimation du maximum de vraisemblance

3.6 Méthodes fondées sur la fonction de vraisemblance

où le paramètre vectoriel α est maintenant choisi pour minimiser un critère de variance scalaire de l'estimateur de variance asymptotique, $\mathcal{A} \operatorname{var}[\hat{\beta}_{mg-q}(a)]$, calculé de la même façon que (3.12).

$$\hat{\beta}_{mg-q}(a) = \left[\sum_{i \in s} q_i a_i a_i' (\alpha) x_i x_i' \right]^{-1} \sum_{i \in s} q_i a_i a_i' (\alpha) x_i y_i. \quad (3.17)$$

les QPDC estimées puissent être utiles pour diverses sortes d'inférences. Voir Pfeffermann (1993) et Binder et Roberts (2009) pour une discussion et des exemples.

L'estimation de la variance sous randomisation des estimateurs à pondération probabiliste est généralement simple en utilisant les techniques disponibles en échantillonnage de population finie. Binder (1983) a élaboré une approche générale pour estimer la variance sous randomisation des estimateurs obtenus comme solution des équations d'estimation pondérées par les probabilités ; voir aussi Binder et Roberts (2009), et Godambe et Thompson (2009). Fuller (1975), Binder (1983), Chamblee et Boyle (1985) et Francisco et Fuller (1991) ont élaboré des théorèmes de la limite centrale applicables aux estimateurs pondérés par les probabilités. Malgré ces propriétés désirables de la pondération probabiliste, la méthode présente certaines limites sérieuses :

1. Elle est limitée principalement à l'estimation ponctuelle. L'inférence probabiliste, telle que les intervalles de confiance ou les tests d'hypothèse, requiert généralement des hypothèses de normalité en grand échantillon. En particulier, la distribution aléatoire ne se prête pas à l'utilisation des méthodes d'inférence classiques, telles que l'inférence fondée sur la fonction de vraisemblance ou l'inférence bayésienne.

2. Les variances des estimateurs à pondération probabiliste sont calculées par rapport à la distribution aléatoire et l'emploi de cette approche ne permet pas le conditionnement sur l'échantillon sélectionné, par exemple, le conditionnement sur les covariables observées ou sur les grappes sélectionnées dans un modèle multiniveaux.
3. Comme il est souvent illustré dans la littérature, les estimateurs à pondération probabiliste ont généralement une plus grande variance que les estimateurs fondés sur un modèle, surtout dans le cas de petits échantillons et d'une forte variation des poids d'échantillonnage.

4. La distribution aléatoire ne permet pas de résoudre les problèmes de prédiction, tels que la valeur au résultat pour les unités non échantillonnées dont les covariables sont connues sous un modèle de régression, ou la prédiction des moyennes de petits domaines pour les domaines sans échantillon dans un problème d'estimation sur petits domaines.

3.5 Modifications des poids d'échantillonnage

Lorsqu'on estime des quantités de population finie, les poids d'échantillonnage sont souvent modifiés en imposant des équations de calage, qui font concorder les estimateurs

PW des covariables pour lesquels les totaux de population sont connus avec les totaux réels. Le recours au calage est particulièrement utile dans le cas de la non-réponse ; voir Kott (2009) pour une discussion récente assortie de références. L'utilisation de la *vraisemblance empirique* dans l'inférence analytique sur des modèles de population, qui tente aussi d'intégrer les équations de calage, quoique d'une façon différente, est discutée plus loin. Ci-après, nous passons en revue deux modifications des poids d'échantillonnage destinées à réduire les variances des estimateurs pondérés des paramètres du modèle sous la *distribution d'échantillon* (2.1). Nous examinons aussi une combinaison des deux modifications.

Magée (1998) examine un modèle de régression linéaire, mais les résultats peuvent être étendus à d'autres modèles de population. L'auteur montre que, sous certaines hypothèses concernant les moments, tout estimateur $\hat{\beta}_{mg}(a) = [\sum_{i \in s} w_i a_i(\alpha) x_i' x_i']^{-1} \sum_{i \in s} w_i a_i(\alpha) x_i' y_i$ ayant des poids positifs $a_i(\alpha) = a(x_i, \alpha)$ est convergent pour β sous la distribution d'échantillon. Les poids $a(x_i, \alpha)$ appartiennent à une famille paramétrisée de fonctions dont le paramètre vectoriel α est choisi pour minimiser un critère de variance scalaire tel que le déterminant ou la trace de l'estimateur de variance asymptotique.

$$A \text{ var}[\hat{\beta}_{mg}(a)]$$

$$= \left[\sum_{i \in s} w_i a_i(\alpha) x_i' x_i' \right]^{-1} \sum_{i \in s} w_i^2 a_i^2(\alpha) \hat{\varepsilon}_i^2 x_i' x_i' \left[\sum_{i \in s} w_i a_i(\alpha) x_i' x_i' \right]^{-1}, \quad (3.12)$$

où $\hat{\varepsilon}_i = (y_i - x_i' \beta^{pw})$. Le choix de la fonction $a(x_i, \alpha)$ est laissé à l'analyste, mais la notion évidente est de choisir une fonction que l'on pense être approximativement inversement proportionnelle à la variance résiduelle sous le modèle d'échantillon. L'auteur montre que l'estimateur « quasi de Aïken » résultant possède asymptotiquement une variance plus faible sous la distribution d'échantillon que l'estimateur à pondération probabiliste $\hat{\beta}^{pw}$. Rappelons que, en vertu de la remarque 8, $\hat{\beta}^{pw}$ est convergent pour β sous la distribution d'échantillon, ce qui justifie de comparer les variances asymptotiques de deux estimateurs sous cette distribution.

Pfeffermann et Sverchokov (1999) proposent une autre modification. Considérons le modèle de population,

$$y_j = m(x_j; \theta) + \varepsilon_j, E^p(\varepsilon_j | x_j) = 0, E^p(\varepsilon_j^2 | x_j) = \sigma^2, \quad (3.13)$$

où $m(x_j; \theta)$ possède une forme connue. Soit $q_i = w_i / E^p(w_i | x_i)$. Les auteurs montrent que, si $\Pr(I_i^l = 1 | \pi_i, y_i, x_i) = \pi_i$,

$$E^p(y_i | x_i) = E^p(w_i y_i | x_i) / E^p(w_i | x_i). \quad (3.14)$$

où ε_{ij} et u_i sont indépendants pour tous i et j . Les paramètres inconnus sont les vecteurs des coefficients $\vartheta = (\beta', \gamma')'$ et les variances $\tau = (\sigma_\varepsilon^2, \sigma_u^2)'$. Supposons que la réponse est complète. Sous échantillonnage ignorable des unités de premier et de deuxième niveau, le *mle* de (ϑ, τ) est calculé de manière commode par itération entre l'estimation de ϑ pour τ « connue » et l'estimation de τ pour ϑ « connu », les valeurs « connues » étant définies par les estimateurs issus de l'itération précédente. Les deux ensembles d'estimateurs sur la r^e itération sont les solutions des équations linéaires de la forme $P^{(r)}(\vartheta) \vartheta = q^{(r)}, R^{(r)}(\tau) = s^{(r)}$, avec définition appropriée des matrices $(P^{(r)}, R^{(r)})$ et des vecteurs $(q^{(r)}, s^{(r)})$, $r = 1, 2, \dots$ (Goldstein 1986). Si on les applique à toutes les valeurs de population, ces équations définissent les équations d'estimation sous recensement.

Supposons, comme auparavant, qu'un échantillon s_1 d'unités de premier niveau est sélectionné avec les probabilités $\pi_i = \Pr(i \in s_1)$, et que des sous-échantillons s_{2i} de taille $n_i < N_i$ sont sélectionnés dans chaque unité de premier niveau sélectionnée i avec les probabilités $\pi_{ji} = \Pr(j \in s_{2i} | i \in s_1)$. Le *pml* pour ce modèle peut être obtenu en exprimant d'abord les éléments des matrices $(P^{(r)}, R^{(r)})$ et des vecteurs $(q^{(r)}, s^{(r)})$ sous forme de sommes sur les unités de premier et de deuxième niveaux, puis en estimant chaque somme de population de la forme $\sum_{i=1}^M d_i'$ au moyen de l'estimateur HT $\sum_{i \in s_1} (d_i' / \pi_i)$, et chaque somme de population de la forme $\sum_{j=1}^{N_i} d_{ji}'$ au moyen de l'estimateur HT $\sum_{j \in s_{2i}} (d_{ji}' / \pi_{ji})$. Voir Pfeffermann, Skinner, Holmes, Goldstein et Rasbash (1998b). Pfeffermann et Sverchkov (2009) passent en revue d'autres méthodes de pondération probabiliste dans les modèles à deux niveaux.

La pondération probabiliste est très répandue pour l'estimation des quantités de populations finie, appelée inférence descriptive dans la littérature, ainsi que pour l'« inférence analytique » sur les modèles de population. Le principal attrait de cette méthode est sa simplicité. Elle est généralement considérée comme « exemple de modèle », sauf lorsque l'on doit estimer les probabilités de réponse, qui sont souvent fondées sur des modèles, et par conséquent plus robuste que les autres méthodes, mais lorsqu'on l'utilise pour l'inférence analytique, ce point de vue est contestable.

Les estimateurs à pondération probabiliste (PW) sont convergents sous randomisation pour les quantités de populations descriptives correspondantes (QPD), définies comme étant les solutions (hypothétiques) des équations d'estimation sous recensement. Cependant, si le modèle de population est mal spécifié, les QPD cibles ne sont pas convergentes sous (le modèle) p pour les paramètres réels du modèle et les estimateurs PW ne sont pas convergents sous $r - p$ non plus. Donc, la pondération probabiliste n'offre aucune protection contre l'erreur de spécification du modèle, quoique

des équations sous recensement (3.6). Sous des conditions générales, $(\theta^{\text{pw}} - \theta^{\text{con}}) = O_p(n^{-0.5})$ et $(\theta^{\text{con}} - \theta) = O_p(N^{-0.5})$, ce qui établit la convergence sous la distribution $r - p$ de θ^{con} dans ces conditions. La variance $r - p$ de θ^{pw} peut être décomposée comme

$$\text{Var}_{r-p}(\theta^{\text{pw}}) = E_p[\text{Var}_r(\theta^{\text{pw}})] + \text{Var}_p[E_r(\theta^{\text{pw}})]. \quad (3.10)$$

Pour l'échantillonnage à un seul degré, si n est beaucoup plus petit que N , comme cela est habituellement le cas, le deuxième terme du deuxième membre de (3.10) est négligable comparativement au premier terme, et $\text{Var}_{r-p}(\theta^{\text{pw}})$ peut être estimé par l'estimateur de variance sous randomisation $\text{Var}_r(\theta^{\text{pw}})$. Ce résultat n'est pas nécessairement vérifié pour l'échantillonnage par grappes, puisque, dans ce cas, $\text{Var}_r(\theta^{\text{pw}})$ est habituellement d'ordre $O(1/m)$, où m est le nombre de grappes échantillonnées, et sous un modèle approprié, $\text{Var}_p[E_r(\theta^{\text{pw}})]$ est d'ordre $O(1/M)$, où M est le nombre de grappes dans la population. Pour que $\text{Var}_p(\theta^{\text{pw}})$ soit un estimateur adéquat de $\text{Var}_{r-p}(\theta^{\text{pw}})$ dans ce cas, m doit être beaucoup plus petit que M .

Remarque 8. La convergence des estimateurs PW sous un modèle de population correctement spécifié peut aussi être établie sous la distribution dans l'échantillon (équation 2.1). Considérons l'estimateur β^{pw} donné par (3.8). Écrivons $\beta^{\text{pw}} = \beta + [\sum_{i \in s} w_i x_i' x_i' / \sum_{i \in s} w_i' w_i' e_i']^{-1} \sum_{i \in s} w_i' w_i' e_i'$, où les e_i sont les résidus du modèle de la population. Le principal résultat menant à la convergence de β^{pw} sous la distribution de l'échantillon est que si $\Pr(I_i = 1 | y_i, x_i, \pi_i) = \pi_i$ alors $E_s(w_i' e_i) = E_s(w_i) E_p(e_i) = 0$, (dérivant de (3.14) plus loin). En fait, en considérant les covariables comme étant aléatoires avec (y_i, x_i) possédant une distribution conjointe,

$$\beta = \arg \min_{\beta} E_p(y_i' - x_i' \beta)^2 = \arg \min_{\beta} E_s[w_i' (y_i' - x_i' \beta)^2],$$

ce qui implique que β^{pw} est l'estimateur optimal (au sens des moindres carrés pondérés) de β sous la distribution d'échantillon de (y_i, x_i) . Voir aussi (3.24) plus loin. Godambe et Thompson (1986, 2009) établissent et discutent d'autres propriétés d'optimalité des estimateurs qui résolvent les équations d'estimation de la forme $\sum_{i \in s} w_i' u_i(y_i')$, $x_i'; \theta) = 0$. L'exemple qui suit montre comment la pondération probabiliste (PW) peut être utilisée pour modéliser les populations en grappes.

Exemple 7. Considérons le modèle (à ordonnée à l'origine aléatoire) à deux niveaux de la population,

$$u_i \sim N(\gamma_i', \sigma_u^2), i = 1, \dots, M$$

Niveau 1 :

$$y_{ij} = x_{ij}' \beta + u_i + \varepsilon_{ij}, \varepsilon_{ij} \sim N(0, \sigma_\varepsilon^2), j = 1, \dots, N_i \quad (3.11)$$

adéquat, on peut omettre de tenir compte de la sélection de l'échantillon pour l'inférence sur les paramètres de $f_p(y_s | x_s, \pi_U)$. Afin d'estimer le modèle cible $f_p(y | x)$ dans ce cas, on peut suivre les mêmes étapes qu'à la section (3.2) en substituant π_U à Z_U .

L'adoption de cette approche réduit la dimension des covariables ajoutées, mais elle requiert que l'on connaisse les probabilités d'inclusion dans l'échantillon (ou les poids d'échantillonnage) pour toutes les unités de la population, information qui pourrait ne pas être disponible dans le cas d'une analyse secondaire. Le cas de la non-réponse pose tout particulièrement problème, puisque les probabilités de réponse sont généralement inconnues et doivent être estimées. Un autre problème important que pose cette approche est que, pour les plans d'échantillonnage généraux, le vecteur π_U pourrait ne pas être un sommaire adéquat de l'information sur le plan nécessaire pour que l'on puisse ignorer l'échantillonnage.

Remarque 5. Même si le vecteur π_U n'est pas toujours un sommaire adéquat de Z_U , pour les plans d'échantillonnage tels que $\Pr(I_i = 1 | y_i, x_i, \pi_i) = \pi_i, f_s(y_i | x_i, \pi_i) = f_p(y_i | x_i, \pi_i)$, de sorte que les f_{dp} marginales de population et d'échantillon, pour une unité d'échantillon donnée, sont néanmoins les mêmes lorsqu'on ajoute π_i aux covariables sont néanmoins les mêmes (voir Skinner 1994).

Remarque 6. Sous les conditions empiriques décrites à la section 3.1, il existe une correspondance de un à un entre les variables du plan d'échantillonnage (z_j^*, z_j^*) et les poids d'échantillonnage (w_j, w_j).

3.4 Méthodes fondées sur la pondération probabiliste

Jusqu'à présent, nous avons considéré des méthodes nécessitant que l'on connaisse les variables J qui déterminent la sélection de l'échantillon et les probabilités de réponse, ou du moins un sommaire adéquat de ces variables. Les méthodes considérées plus bas requièrent seulement la connaissance des poids d'échantillonnage pour les unités échantillonnées qui répondent. En ce sens, elles sont limitées aux situations de réponse complète, ou aux cas où les probabilités de réponse peuvent être estimées avec suffisamment de précision, auquel cas le poids d'échantillonage d'une unité répondante est égal à l'inverse du produit de la probabilité de sélection de l'unité et de sa probabilité de réponse estimée. La pondération probabiliste (PP) est discutée dans de nombreux articles ; voir la discussion récente dans Pfeffermann et Sverchkov (2009) et les références incluses. Comme auparavant, nous nous concentrons ici sur l'estimation de modèles de population.

Pour présenter l'idée, considérons le cas d'un recensement dans lequel la réponse est complète. En supposant que les résultats sont indépendants, les paramètres du modèle, θ , sont habituellement estimés dans ce cas en résolvant les équations d'estimation sous recensement de la forme,

$$\sum_{j=1}^N n(y_j, x_j; \theta) = 0. \quad (3.6)$$

Dans le cas du *mle*, $n(y_j, x_j; \theta) = (\partial/\partial\theta) \log f_p(y_j | x_j; \theta)$, le f^* score. En pratique, les données ne sont disponibles que pour un échantillon $s \subset U$ et les équations (3.6) sont remplacées par leur estimateur de Horvitz-Thompson sans biais sous randomisation,

$$\sum_{i \in s} w_i n(y_i, x_i; \theta) = 0, \quad (3.7)$$

où les w_i sont les poids d'échantillonnage.

Remarque 7. Quand les équations d'estimation sous recensement (3.6) sont les équations de vraisemblance, les estimateurs obtenus en résolvant (3.7) sont appelés dans la littérature sur l'échantillonnage « estimateur du pseudo-maximum de vraisemblance » (*pml*). Voir Binder (1983), Skinner et coll. (1989), Pfeffermann (1993, 1996), et Godambe et Thompson (2009) pour une discussion assortie de nombreux exemples. Cette approche est implémentée dans de nombreux logiciels, tels que SAS, STATA ou SUDAAN.

Exemple 5. Dans le cas du modèle de régression linéaire classique, le *pml* ou l'estimateur à pondération probabiliste (PW, pour *probability weighting*) du vecteur de coefficients β résout les équations $\sum_{i \in s} w_i (y_i - x_i' \beta^{pw}) x_i = 0$;

$$\hat{\beta}^{pw} = \left[\sum_{i \in s} w_i x_i x_i' \right]^{-1} \sum_{i \in s} w_i x_i y_i. \quad (3.8)$$

L'estimateur PW de la variance résiduelle est $\hat{\sigma}^2_{pw} = \sum_{i \in s} w_i (y_i - x_i' \hat{\beta}^{pw})^2 / (\sum_{i \in s} w_i - k)$, où $k = \dim(\beta)$. Dans le cas de la régression logistique, les équations de pseudo-vraisemblance (sans aucune solution explicite) sont,

$$\sum_{i \in s} w_i [y_i - \tilde{p}_i(x_i)] x_i = 0 ; \tilde{p}_i(x_i)$$

$$= \Pr(y_i = 1 | x_i)$$

$$= \exp(x_i' \beta) / [1 + \exp(x_i' \beta)]. \quad (3.9)$$

Exemple 6. Soit $n(y_j; \theta) = [\Delta(\theta - y_j) - F_p(\theta)]$, où $F_p(\theta)$ est la fonction de répartition de la population à θ , et $\Delta(a) = 1(0) \leq a < 0$. L'estimateur PW de $F_p(\theta)$ est $F_{p,pw}(\theta) = \sum_{i \in s} w_i \Delta(\theta - y_i) / \sum_{i \in s} w_i$, l'estimateur bien connu de Hájek (1971).

La propriété des estimateurs PW qui mérite d'être mentionnée est qu'ils sont généralement convergents sous $r - p$. Voir la section 2.2 pour la définition de la distribution $r - p$. On peut le voir en décomposant $(\hat{\theta}^{pw} - \theta) = (\hat{\theta}^{pw} - \hat{\theta}^{cen}) + (\hat{\theta}^{cen} - \theta)$, où $\hat{\theta}^{cen}$ est la solution (hypothétique)

L'utilisation de l'approche est intéressante et a l'avantage de permettre d'utiliser les procédures classiques d'inférence fondées sur un modèle une fois que les variables $J_U = Z_U \cup L_U$ sont incluses dans le modèle, mais elle est souvent limitée en pratique pour les raisons suivantes :

1. Elle requiert que l'on connaisse les valeurs de population de toutes les variables déterminant la sélection de l'échantillon et la réponse, et cette information est habituellement inconnue de l'analyste qui ajuste le modèle à cause des contraintes de confidentialité ou d'autres raisons. Même si elles sont connues, inclure dans le modèle toutes les variables géographiques et opérationnelles utilisées pour établir le plan d'échantillonnage et les variables expliquant la réponse peut être une tâche énorme.
2. En pratique, il peut exister de nombreuses covariables et de nombreuses variables du plan d'échantillonnage, et la modélisation de la relation entre les premières et les secondes afin d'éliminer par intégration l'effet des premières peut être compliquée et ne plus reproduire le modèle cible original.

Feder (2011) propose la solution simple qui suit pour ce problème. Supposons d'abord que les variables du plan d'échantillonnage et les covariables sont connues pour chaque élément de la population. La solution proposée consiste à imputer les résultats manquants dans la population en utilisant le modèle $f^d(y_s | x_s, J_U = J_U)$ ajusté aux données d'échantillon, puis d'ajuster le modèle de population $f^p(y_j | x_j)$ en utilisant toutes les valeurs de population, les résultats manquants étant remplacés par leur valeur imputée. Quand les variables du plan d'échantillonnage et les covariables sont inconnues pour les unités non échantillonnées, elles doivent être imputées également. L'imputation peut être effectuée par échantillonnage avec remise de $(N - n)$ valeurs (x_i, z_i) parmi les valeurs d'échantillon avec les probabilités $\bar{p}_i = (w_i - 1) / \sum_{k=1}^n (w_k - 1)$ à chaque tirage, où les w_i sont les poids d'échantillon. Voir Pfeffermann et Sikov (2011) pour la justification de cette procédure sous le modèle d'échantillon et l'extension du cas à la non-réponse de type NMAR.

3. L'approche n'est pas applicable quand l'inclusion dans l'échantillon dépend aussi des valeurs résultat, c'est-à-dire $Z_U = \{Y_U, Z_U^*\}$ et $\Pr(A_s = 1 | Y_U, X_U, Z_U^*) \neq \Pr(A_s = 1 | X_U, Z_U^*)$. Un exemple classique est celui des *études cas-témoins* (Scott et Wild 2009), mais un problème semblable se pose quand la non-réponse est de type NMAR.

Remarque 4. Inclure les variables du plan d'échantillonnage et les variables expliquant la réponse dans le modèle ne nécessite pas forcément de les éliminer par intégration,

Exemple 4 : Supposons qu'un échantillon de taille n est sélectionné avec les probabilités définies par les valeurs de population des variables du plan d'échantillonnage Z et que toutes les unités échantillonnées répondent. Supposons que la distribution de population de Y, X, Z suit une loi normale multivariée. Les données dont dispose l'analyste sont les valeurs d'échantillon $[y_s, x_s]$ et les valeurs de population Z_U . En utilisant les propriétés de la loi normale multivariée, $E^p(Y | x) = \beta_0 + \beta_{yx}x$ pour certains coefficients (β_0, β_{yx}) , mais l'estimation par les moindres carrés ordinaires de β_{yx} est biaisée, parce que les probabilités d'échantillonnage dépendent de Z , qui est corréliée à Y . Le *mle* de β_{yx} pour le cas d'une loi normale trivariée est (DeMets et Halperin 1977),

$$\hat{\beta}_{yx} = \left\{ s_{xy} + \frac{s_{yz}s_{xz}}{s_{zz}} \left(\frac{\hat{\sigma}_z^2}{s_{zz}} - 1 \right) \right\} / \left\{ s_{xx} + \frac{s_{xz}^2}{s_{zz}} \left(\frac{\hat{\sigma}_z^2}{s_{zz}} - 1 \right) \right\}, \quad (3.5)$$

où $s_{uv} = n^{-1} \sum_{i=1}^n (u_i - \bar{u})(v_i - \bar{v})$ et $\hat{\sigma}_z^2 = N^{-1} \sum_{i=1}^N (z_i - \bar{z}_U)^2$, avec \bar{u}, \bar{v} et \bar{z}_U définissant les moyennes d'échantillon et de population correspondantes. Donc, les valeurs de population de Z figurent dans ce cas dans l'estimateur optimal du paramètre cible β_{yx} . Holt et coll. (1980) étendent ce résultat au cas où Y, X, Z sont des variables vectorielles. Nathan et Holt (1980) établissent les conditions sous lesquelles $\hat{\beta}_{yx}$ est convergent sans les hypothèses de normalité multivariée. Pfeffermann et Holmes (1985) étudient la robustesse de l'estimateur à l'erreur de spécification du modèle.

3.3 Utilisation des poids d'échantillonnage comme substituts des variables du plan d'échantillonnage

Dans les situations où les variables du plan d'échantillonnage déterminant la sélection de l'échantillon sont trop nombreuses pour les introduire toutes dans le modèle, ou quand certaines de ces variables ou toutes sont inconnues de l'analyste, il est souvent recommandé d'inclure dans le modèle les poids d'échantillonnage comme substituts des variables du plan. Des exemples de l'utilisation de cette approche sont décrits dans DuMouchel et Duncan (1983), Samdal et Wright (1984), Rubin (1985), Chambers, Dorfman et Wang (1998), et Wu et Fuller (2006). Rubin (1985) définit le vecteur $a = (a_1, \dots, a_N)' = a(Z_U)$ de façon qu'il soit un sommaire adéquat de Z_U si $\Pr(A_s = 1 | Z_U) = \Pr(A_s = 1 | a)$. Il montre que le vecteur $\pi_U = (\pi_1, \dots, \pi_N)$ des probabilités d'inclusion dans l'échantillon est le sommaire adéquat possible le plus grossier de Z_U , bien qu'il puisse être trop grossier. Il s'ensuit donc que, pour des plans d'échantillonnage tels que $\Pr(A_s = 1 | X_U, Z_U) = \Pr(A_s = 1 | Z_U)$, si π_U est un sommaire

La *fdp* (2.11) couplée à la *fap* de niveau 1 donnée par (2.10) définit le modèle vérifié pour les données observées dans le cas d'un échantillonnage en grappes informatif et d'une non-réponse de type NMAR.

3. Comment estimer les modèles de population d'après des données d'enquêtes complexes ?

Dans la présente section, nous passons en revue les principales approches proposées dans la littérature pour traiter les caractéristiques particulières des données d'enquêtes complexes exposées à la section 2 et proposons certaines modifications. Afin de simplifier la discussion, nous tenons compte des conditions qui suivent, utilisées pour l'étude en simulation décrite à la section 4.

3.1 Modèle de population et plan d'échantillonnage

Considérons une population stratifiée $U = U_1 \cup \dots \cup U_H$ de taille N . Spécifiquement, définissons pour chaque unité $j \in U$ un indicateur de stratification vectoriel aléatoire $z_j = (z_{1j}, \dots, z_{Hj})'$ tel que $\Pr(z_{hj} = 1) = p_h$, $\sum_{h=1}^H p_h = 1$ et $j \in U_h$ si $z_{hj} = 1$. La stratification est effectuée indépendamment d'une unité à l'autre. Les valeurs d'une variable de résultat Y sont produites comme étant $y_j = \beta_0 + \beta_1 x_j + \alpha_0 \zeta_j + \alpha_1 \zeta_j x_j + \varepsilon_j$; $\varepsilon_j \sim N(0, \sigma^2)$, où les x_j sont des covariables scalaires fixes, les $(\beta_0, \beta_1, \alpha_0, \alpha_1)$ sont des coefficients fixes et

$$\zeta_j = \frac{1}{\sum_{h=1}^H \frac{p_h}{z_{hj}}} - 1.$$

Notons que ζ_j est une variable aléatoire de moyenne nulle et de variance

$$V_{\zeta} = \left(\frac{1}{\sum_{h=1}^H \frac{H}{p_h}} - 1 \right),$$

ce qui implique que, pour les covariables données x_j, x_k ,

$$\begin{aligned} E_p(y_j | x_j) &= \beta_0 + \beta_1 x_j, \text{Var}_p(y_j | x_j) \\ &= (\alpha_0 + \alpha_1 x_j)^2 V_{\zeta} + \sigma^2, \text{Cov}_p(y_j, y_k | x_j, x_k) \\ &= 0, j \neq k. \end{aligned} \quad (3.1)$$

Cependant, pour l'unité $j \in U_h$

$$y_j | x_j, z_{hj} = 1 \sim N[(\beta_0 + \alpha_0 \zeta_h)$$

$$+ (\beta_1 + \alpha_1 \zeta_h) x_j, \sigma^2]; \zeta_h = [(1 / H p_h) - 1]. \quad (3.2)$$

Donc, dans chaque strate, le modèle de régression est le modèle linéaire classique avec variance constante, mais l'ordonnée à l'origine et la pente changent d'une strate à l'autre.

Le modèle défini par (3.1) et (3.2) est un modèle de régression à coefficients aléatoires raisonnable qui, selon nous, reproduit un grand nombre de populations rencontrées en pratique.

Nous avons utilisé l'échantillonnage systématique avec probabilité proportionnelle à la taille (PPT) pour tirer les échantillons dans les strates en définissant la variable de strate comme étant $z_j^* = \max\{\min\{[q_j]^{1/5}, 9\}, 1\}$; $q_j \sim N(1 + x_j, 1)$. Le choix de cette variable de strate n'a rien de nouveau, excepté qu'il permet de faire clairement la distinction entre la variance des divers estimateurs. Cette taille z_j^* ne dépend pas du résultat y_j , et donc le processus d'échantillonnage dans chaque strate est non informatif. Toutefois, en cas de répartition non proportionnelle de l'échantillon entre les strates, le plan d'échantillonnage est informatif en raison des différents modèles appliqués dans les diverses strates, si bien que les résultats observés contiennent de l'information sur l'appartenance à la strate et $\Pr(j \in s | y_j, x_j) \neq \Pr(j \in s | x_j)$. Nous nous concentrons sur l'estimation des coefficients de régression (β_0, β_1) dans (3.1) comme cible de l'inférence et supposons que l'information sur l'échantillon disponible comprend les résultats observés et les covariables, les vecteurs d'appartenance aux strates z_{hj} et les tailles de strates $\{N_h\}$.

3.2 Ajout des variables du plan d'échantillonnage aux covariables

Comme l'implique (2.3), le modèle de la population $(f_{dp}, f_{dp}^p | x_s)$ et le modèle de l'échantillon $f_s(y_s | x_s)$ sont identiques si $\Pr(A_s = 1 | y_s, x_s) = \Pr(A_s = 1 | x_s) \forall y_s$. En vertu de (2.2), le processus de réponse est ignorable si $\Pr(R_i = 1 | y_i, x_i, I_i = 1) = \Pr(R_i = 1 | x_i, I_i = 1) \forall y_i$. Donc, un moyen possible de tenir compte des effets de l'échantillonnage et de la réponse est d'ajouter aux covariables des modèles toutes les variables et interactions déterminant les probabilités d'échantillonnage et de réponse, puis à les intégrer afin d'estimer le modèle d'intérêt. Désignons ces variables par $J = Z \cup L$ avec les valeurs de population J_U , où L définit les variables expliquant les probabilités de réponse. En supposant que $f_{dp}^p(y_s | x_U, J_U) = f_{dp}^p(y_s | x_s, J_U)$, l'utilisation de cette approche requiert d'ajuster d'abord le modèle

$$f_{dp}^p(y_s | x_s, J_U = j_U) = \int f_{dp}(y_s, y_s | x_U, J_U) dy_s, \quad (3.3)$$

et puis d'intégrer,

$$f_{dp}^p(y_s | x_s) = \int f_{dp}(y_s | x_s, j_U) f(j_U | x_s) dj_U. \quad (3.4)$$

Des variantes de cette approche sont décrites dans DeMets et Halperin (1977), Holt, Smith et Winter (1980), Nahan et Holt (1980), Jewell (1985), Skinner (1994), Chambers et Skinner (2003, chapitre 2) et Gelman (2007).

$\theta_2 = (\beta, \sigma_e^2)$. Supposons que l'échantillon est tiré selon le processus d'échantillonnage à deux degrés suivant. Au premier degré, un échantillon s_1 de $m < M$ unités de premier niveau (grappes ; disons les écoles) est sélectionné avec les probabilités $\pi_i = \Pr(i \in s_1)$ qui peuvent être corrélées aux effets aléatoires u_i après conditionnement sur les covariables t_i . Au deuxième degré, un sous-échantillon s_{2i} de $n_i < N_i$ unités de deuxième niveau (unités finales d'échantillonnage ; disons les élèves) est tiré de chaque unité de premier niveau sélectionnée i avec les probabilités $\pi_{ji} = \Pr(j \in s_{2i} | i \in s_1)$ qui peuvent être corrélées aux résultats y_{ji} après conditionnement sur les covariables x_{ji} . Désignons par I_i et I_{ji} les indicateurs d'échantillonnage de premier et de deuxième degrés. En vertu de (2.1), le modèle d'échantillonnage à deux niveaux vérifié pour les données observées, correspondant au modèle de population (2.9), est

$$\begin{aligned} \text{Niveau 1 :} \\ f_{s_1}(u_i | t_i; \theta_1, \gamma_1) &= \frac{\Pr(I_i = 1 | u_i, t_i; \gamma_1) \phi^p(u_i | t_i; \theta_1)}{\Pr(I_i = 1 | t_i; \theta_1, \gamma_1)} \\ \text{Niveau 2 :} \\ f_{s_{2i}}(y_{ji} | x_{ji}, u_i; \theta_2, \gamma_2) &= \frac{\Pr(I_{ji} = 1 | y_{ji}, x_{ji}; \gamma_2) f^p(y_{ji} | x_{ji}, u_i; \theta_2)}{\Pr(I_{ji} = 1 | x_{ji}, u_i; \theta_2)} \end{aligned} \quad (2.10)$$

où nous supposons que $\Pr(I_{ji} = 1 | y_{ji}, u_i, x_{ji}; \gamma_2) = \Pr(I_{ji} = 1 | y_{ji}, x_{ji}; \gamma_2)$.

Remarque 3. En vertu du résultat d'indépendance de la remarque 2, si les $y_{ji} | u_i$ sont indépendants sous le modèle de population, ils sont asymptotiquement indépendants sous le modèle d'échantillon. De même, si les effets aléatoires u_i sont indépendants sous le modèle de population, ils sont asymptotiquement indépendants sous le modèle d'échantillon (2.10) est un vrai modèle à deux niveaux, quoiqu'avec des distributions différentes et peut-être un plus grand nombre de paramètres. Evidemment, les modèles (2.9) et (2.10) sont différents, à moins que $\Pr(I_{ji} = 1 | y_{ji}, x_{ji}) = \Pr(I_{ji} = 1 | u_i, x_{ji})$ et $\Pr(I_i = 1 | u_i, t_i) = \Pr(I_i = 1 | t_i)$.

Jusqu'à présent, j'ai implicitement émis l'hypothèse que la réponse était complète. Supposons, par exemple, que dans la grappe échantillonnée (unité de premier niveau) i seul un sous-échantillon $r_{2i} \subset s_{2i}$ répond, et désignons par R_{ji} l'indicateur de réponse. Le modèle de deuxième niveau pour les résultats observés devient maintenant

$$\begin{aligned} f_{s_{2i}}(y_{ji} | x_{ji}, u_i; \theta_2, \gamma_2, \gamma_2^*) &= f(y_{ji} | x_{ji}, u_i; \theta_2, \gamma_2, \gamma_2^*) \\ &= f(y_{ji} | x_{ji}, u_i; \theta_2, \gamma_2, \gamma_2^*) \\ &= \frac{\Pr(R_{ji} = 1 | y_{ji}, x_{ji}, u_i; \theta_2, \gamma_2, \gamma_2^*)}{\Pr(R_{ji} = 1 | y_{ji}, x_{ji}, u_i; \theta_2, \gamma_2, \gamma_2^*)} \cdot \Pr(R_{ji} = 1 | x_{ji}, u_i; \theta_2, \gamma_2, \gamma_2^*) \end{aligned} \quad (2.11)$$

grappes, due à l'utilisation d'échantillons en grappes à plusieurs degrés. Les grappes sont des « groupes naturels » tels que les ménages, les îlots de résidences, les écoles, voire même les individus dans le cas d'enquêtes longitudinales. Par conséquent, il existe généralement entre les résultats concernant une même grappe une corrélation appelée *corrélation intra-classe*. Il est important de souligner que les grappes représentent un groupement de population existant, de sorte qu'une corrélation intra-classe existe aussi sous le modèle de population.

Pfeffermann et Smith (1985) passent en revue plusieurs classes de modèles de régression plausibles pour les populations en grappes, et discutent des moyens de les estimer à partir de l'échantillon. Un modèle de population utilise fréquemment est le modèle à ordonnée à l'origine aléatoire,

$$y_{ji} = x_{ji}'\beta + u_i + \varepsilon_{ji}; \quad i = 1, \dots, M, \quad j = 1, \dots, N_i; \quad \text{indép.} \quad (2.7)$$

où M définit le nombre de grappes et N_i le nombre d'unités dans la grappe i . Le modèle suppose aussi que $E(u_i \varepsilon_{ji}) = 0, \forall i, j$. Soulignons que, sous ce modèle, $\text{Var}(y_{ji}) = \sigma_u^2 + \sigma_e^2, E(y_{ji} y_{ji}') = \sigma_u^2$ pour $j \neq l$ et $E(y_{ji} y_{li}') = 0$ pour $i \neq k$, ce qui implique

$$\begin{aligned} \text{Corr}(y_{ji}, y_{li}) &= \sigma_e^2 / (\sigma_u^2 + \sigma_e^2) \quad \text{pour } j \neq l; \\ \text{Corr}(y_{ji}, y_{li}) &= 0 \quad \text{pour } i \neq k. \end{aligned} \quad (2.8)$$

Scott et Holt (1982) montrent que l'estimation de β dans (2.7) par les moindres carrés ordinaires (OLS, pour *ordinary least squares*) aboutit habituellement à une faible perte d'efficacité comparativement à l'utilisation de l'estimateur optimal par les moindres carrés généralisés. Toutefois, ne pas tenir compte de la corrélation intra-grappe lorsque l'on estime la variance de l'estimateur OLS peut sous-estimer considérablement cette dernière et donc fausser la puissance des statistiques de test et produire des intervalles de confiance trop courts.

Les résultats présentés dans Scott et Holt (1982) et dans Pfeffermann et Smith (1985) reposent sur l'hypothèse que l'échantillonnage est non informatif et que la réponse est complète. S'il n'en est pas ainsi, le modèle vérifié pour les données d'échantillon diffère du modèle de population correspondant, mais les grappes du modèle sont préservées comme nous allons le montrer. Considérons le modèle de population à deux niveaux suivant :

$$\begin{aligned} \text{Niveau 1 : } u_i | t_i &\sim \phi^p(u_i | t_i; \theta_1), \quad i = 1, \dots, M \\ \text{Niveau 2 : } y_{ji} | (u_i, x_{ji}) &\sim f^p(y_{ji} | x_{ji}, u_i; \theta_2), \quad j = 1, \dots, N_i \end{aligned} \quad (2.9)$$

où ϕ^p et f^p désignent les *fdp* de premier et de deuxième niveau avec les covariables connues t_i et x_{ji} , régies par les hyperparamètres θ_1 et θ_2 , respectivement. Le modèle (2.7) est un cas particulier de (2.9) dans lequel ϕ^p et f^p sont des *fdp* normales avec $t_i = 0$ (pas de covariables), $\theta_1 = \sigma_u^2$ et $\theta_2 = \sigma_e^2$.

condition clé pour pouvoir ignorer le processus d'échantillonnage étant donné l'information sur le plan est que $A_s \perp Z_U | d_s$, avec \perp signifiant l'indépendance ce qui implique que $\Pr(A_s = 1 | Z_U = z_U) = \Pr(A_s = 1 | d_s)$ pour toute valeur z_U pour laquelle $D_s(z_U) = d_s$.

Dans le cas des enquêtes par sondage à plusieurs degrés à grande échelle pour lesquelles les variables du plan d'échantillonnage peuvent être nombreuses, il est généralement difficile et souvent peu pratique de vérifier directement les conditions qui permettent de ne pas tenir compte de la sélection de l'échantillon ou de la non-réponse étant donné l'information sur le plan d'échantillonnage disponible. Par ailleurs, même quand la *fdp* d'échantillon diffère de la *fdp* de population, l'inférence en ignorant le processus d'échantillonnage n'est pas nécessairement fausse. Pour l'illustrer simplement, considérons le cas particulier de l'exemple 1 où $\gamma_2 = 0$. Dans ce cas, la *fdp* d'échantillon est normale et possède les mêmes coefficients de pente et variance résiduelle que la *fdp* de population. Donc, pour l'inférence au sujet des coefficients de pente, on peut ignorer le processus d'échantillonnage. Un résultat similaire tient pour les modèles logistiques quand la sélection de l'échantillon dépend de y mais non de x . Voir Pfeffermann et coll. (1998a) pour le calcul de ce résultat. Pfeffermann et Sverchkov (2009) passent en revue plusieurs statistiques de test proposées dans la littérature pour déterminer si le fait de ne pas tenir compte de la sélection de l'échantillon est justifié pour l'inférence voulue.

2.2 L'utilisation de la distribution aléatoire pour l'inférence

Une caractéristique unique des enquêtes par sondage est que l'échantillon est sélectionné au hasard conformément à un plan d'échantillonnage $\{s, \Pr(s)\}$, $s \in S$. Le plan d'échantillonnage induit une *distribution aléatoire* (discrète) pour toute statistique T_{ys} , qui est la distribution conditionnelle sur toutes les sélections possibles de l'échantillon, sachant les valeurs de population finie. Donc, la statistique T_{ys} prend la valeur t_{ys} avec la probabilité $\Pr(s)$, $s \in S$. L'inférence fondée sur l'échantillonnage classique repose uniquement sur cette distribution. Par exemple, l'estimateur de Horvitz-Thompson (HT) bien connu T_{HT}^{ys} , qui prend la valeur $t_{HT}^{ys} = \sum_{i \in s} y_i / \pi_i$ si l'échantillon s est tiré, est sans biais sous randomisation pour le total de population finie $TOT^{ys} = \sum_{i \in N} y_i$, puisque $\sum_{s \in S} \Pr(s) t_{HT}^{ys} = T^{ys}$. Sa variance est $\text{Var}(T_{HT}^{ys}) = \sum_{s \in S} \Pr(s) (t_{HT}^{ys} - T^{ys})^2$. Notons qu'en cas de non-réponse, l'utilisation de la distribution aléatoire requiert la connaissance des probabilités de réponse, qui ne peuvent être qu'estimées en pratique. L'estimateur HT prend dans ce cas la forme $T_{HT}^{ys} = \sum_{i \in R} y_i / [\pi_i \times \Pr(R_i = 1 | I_i = 1)]$, où R définit le sous-échantillon de répondants. Voir Fuller (2002) pour une discussion plus approfondie.

La distribution aléatoire est conditionnée par les valeurs de population réalisées. Par conséquent, elle peut être utilisée pour l'inférence descriptive sur des fonctions connues des valeurs de population finie, mais non pour l'inférence analytique sur un modèle hypothétique produisant ces valeurs. Pour cela, on pourrait considérer la distribution conjointe sur tous les résultats d'échantillon possibles pour les valeurs de population données (la distribution r sous *randomisation*) et toutes les réalisations possibles des mesures de population finie (la distribution p sous un *modèle*). Voir Binder et Roberts (2009) et les références incluses. La distribution conjointe $r - p$ offre un autre cadre d'inférence pour l'utilisation des *fdp* $f_s(y|x)$ ou $f_o(y|x)$ définies plus haut.

Exemple 3 : Supposons que le modèle de population est $y_i \sim \text{Mult}\{p_k, K\}$, tel que $\Pr_p(y_i = k) = p_k$, $k = 1, \dots, K$; $\sum_{k=1}^K p_k = 1$. Soit $\Pr(i \in s | y_i = k) = \pi_k$. Alors, en vertu de (2.1), $\Pr_s(y_i = k) = \Pr(y_i = k | i \in s) = \pi_k p_k / \sum_{k=1}^K \pi_k p_k = p_k^*$, ou $y_i | i \in s \sim \text{Mult}\{p_k^*, K\}$. En supposant que les résultats observés et les probabilités de sélection connues sont indépendants, l'estimateur du maximum de vraisemblance (*mle*, pour *maximum likelihood estimator*) de p_k fondé sur la distribution dans l'échantillon est $\hat{p}_k = (n_k / \pi_k) / \sum_{j=1}^K (n_j / \pi_j)$, où n_k est le nombre d'unités échantillonnées présentant le résultat $y_i = k$. L'utilisation de la distribution $r - p$ suggère que l'on estime p_k par l'estimateur HT $\hat{p}_k = (1/N) \sum_{i|y_i=k} (1/\pi_k) = (n_k / \pi_k) / N$. L'estimateur \hat{p}_k est sans biais sous la randomisation r pour \hat{p}_k est sans biais sous le modèle p pour p_k , de sorte que \hat{p}_k est sans biais sous

La différence évidente entre la distribution $r - p$ et la distribution d'échantillon, $f_s(y|x)$, est que la seconde est conditionnée sur l'échantillon observé d'unités (et donc les valeurs observées des covariables ou des grappes sélectionnées dans un échantillon en grappes), tandis que la distribution $r - p$ tient compte de toutes les sélections d'échantillons possibles. Par conséquent, l'utilisation de cette dernière distribution ne se prête généralement pas à l'inférence conditionnelle. L'utilisation des *fdp* $f_s(y|x)$ ou $f_o(y|x)$ requiert la modélisation de $\Pr(I_i = 1 | x_i, y_i)$ (équation 2.1) et de $\Pr(R_i = 1 | y_i, x_i, I_i = 1)$ en cas de non-réponse (équation 2.2), mais elle permet le calcul (l'estimation) de la *fdp* conditionnelle des résultats observés sachant les covariables et, donc, l'utilisation des outils d'inférence classiques.

2.3 Données obtenues à partir d'un échantillon en grappes

Une autre caractéristique particulière des données d'enquête mentionnée dans l'introduction est la mise en

indépendantes, les mesures d'échantillon sont *asymptotiquement indépendantes* sous la distribution d'échantillon. Le cadre asymptotique requiert que la taille de la population augmente, mais que la taille de l'échantillon demeure fixe. Comme il est illustré à la section 2.3, il arrive souvent que l'hypothèse des mesures de population indépendantes ne soit pas contraignante non plus.

Jusqu'à présent, par souci de commodité, nous avons supprimé de la notation les paramètres qui sous-tendent la *fdp* de population et le processus d'échantillonnage. Considérons, par exemple, la *fdp* d'échantillon (2.3). En ajoutant la notation des paramètres, elle peut s'écrire

$$f_s(y_s | x_s; \theta, \gamma) = \frac{\Pr(A_s = 1 | y_s, x_s; \gamma) f_p(y_s | x_s; \theta)}{\Pr(A_s = 1 | x_s; \theta)}. \quad (2.5)$$

Donc, les *fdp* conditionnelles de population et d'échantillon sont différentes, à moins que

$$\Pr(A_s = 1 | y_s, x_s; \gamma) = \Pr(A_s = 1 | x_s; \theta, \gamma) \quad \forall y_s. \quad (2.6)$$

Quand l'expression (2.6) est vérifiée, l'inférence sur le paramètre cible θ peut être effectuée en ajustant le modèle de population sur les données d'échantillon, sans tenir compte de la sélection de l'échantillon. Notons que cette conclusion se rapporte à l'échantillon sélectionné défini par l'événement $A_s = 1$.

La condition (2.6) est forte. Dans un article fondamental sur les valeurs manquantes, Rubin (1976) établit les conditions sous lesquelles le processus d'échantillonnage peut être ignoré pour l'inférence fondée sur la fonction de vraisemblance, sur un modèle bayésien ou sur la théorie d'échantillonnage (échantillonnage répété à partir d'un modèle), c'est-à-dire des conditions sous lesquelles le modèle de population défini par $f_p(y_s | x_s; \theta)$ peut être ajusté aux données observées en fonction de la méthode d'inférence. Little (1982) étend les résultats de Rubin en faisant la distinction entre la sélection de l'échantillon et le processus de réponse. Une autre distinction importante tient au fait que Little conditionne sur les valeurs de population Z_U des variables du plan utilisées pour la sélection de l'échantillon. L'inférence sur le modèle de population cible $f_p(y_s | x_s; \theta)$ requiert par conséquent l'intégration de la *fdp* conditionnelle de $y_s | Z_U, x_s$ sur la distribution de $Z_U | x_s$ (voir la section 3). Sugden et Smith (1984) ont établi les conditions sous lesquelles un processus d'échantillonnage qui dépend des variables du plan Z est ignorable, étant donné l'information partielle sur le plan d'échantillonnage. Posons que $d_s = D_s(z_U)$ contient toute l'information disponible sur le plan tel que la strate d'appartenance (qui peut n'être connue que pour les unités échantillonnées), les probabilités d'échantillonnage *etc.* En utilisant la notation antérieure, une

de sorte que les résultats d'échantillon sont aussi indépendants.

$$= \prod_{i \in s} f_s(y_i | x_i), \quad (2.4)$$

$$f_s(y_s | x_s) = \prod_{i \in s} \frac{\Pr(I_i = 1 | y_i, x_i) f_p(y_i | x_i)}{\Pr(I_i = 1 | x_i)}$$

La *fdp* $f_p(y_s | x_s)$ peut être générale, permettant en particulier des mesures corrélées, mais la modélisation de la probabilité $\Pr(A_s = 1 | y_s, x_s)$ n'est pratiquement faisable que si l'échantillon peut être décomposé en sous-ensembles exclusifs et exhaustifs s_k tels que $\Pr(A_s = 1 | y_s, x_s) \propto \prod_k \Pr(A_{s_k} = 1 | y_{s_k}, x_{s_k})$, et que $\Pr(A_s = 1 | y_{s_k}, x_{s_k})$ satisfasse le même modèle pour tous les sous-ensembles (voir l'exemple 2). En particulier, si les résultats de population sont indépendants sachant les covariables sous le modèle de population et que $\Pr(A_s = 1 | y_s, x_s) \propto \prod_{i \in s} \Pr(I_i = 1 | y_i, x_i)$, (2.3) prend la forme

$$= \frac{\Pr(A_s = 1 | y_s, x_s) f_p(y_s | x_s)}{\Pr(A_s = 1 | x_s)}. \quad (2.3)$$

$$f_s(y_s | x_s) = f(y_s | x_s, A_s = 1)$$

Exemple 2. Considérons le cas d'une population en grappes $U = \bigcup_i U_i$ et de mesures indépendantes entre les grappes, telles que $f_p(y_U | x_U) = \prod_i f_p(y_{U_i} | x_{U_i})$, où (y_U, x_U) définit toutes les valeurs de population et (y_{U_i}, x_{U_i}) , les valeurs dans la grappe i . Soit s définissant l'ensemble des grappes échantillonnées de manière indépendante avec probabilités $\Pr(i \in s | y_U, x_U) = r(y_{U_i}, x_{U_i})$ pour une fonction quelconque $r(\cdot)$, et supposons aussi que toutes les unités dans les grappes échantillonnées sont observées (échantillonnage en grappes à un degré). Alors, $\Pr(A_s = 1 | y_U, x_U) = \prod_{k \in s} r(y_{U_k}, x_{U_k}) \times \prod_{j \notin s} [1 - r(y_{U_j}, x_{U_j})]$. Puisque, pour $k \in s$, $(y_{U_k}, x_{U_k}) = (y_{s_k}, x_{s_k})$, il s'ensuit que $\Pr(A_s = 1 | y_s, x_s) = \prod_{k \in s} r(y_{s_k}, x_{s_k}) \times G$, où pour les covariables données $x_{U_j}, j \notin s$, G est une constante qui satisfait $G = \prod_{j \notin s} [1 - r(y_{U_j}, x_{U_j})]$. Une population non en grappes avec des mesures indépendantes et un échantillonnage de Poisson des unités individuelles est un cas particulier où chaque grappe est constituée d'un seul élément, ce qui mène à (2.4).

Remarque 2. Les exemples considérés jusqu'à présent supposent un échantillonnage indépendant, qui préserve l'indépendance des résultats après l'échantillonnage, mais cette hypothèse peut habituellement être relâchée en suivant un résultat prouvé et illustré dans Pfeffermann et coll. (1998). En vertu de ce résultat, sous certaines conditions de régularité générale et pour de nombreux scénarios d'échantillonnage fréquemment utilisés pour la sélection avec probabilités inégales, si les mesures de population sont

souvent que les deux ensembles de covariables ne soient pas identiques. Cependant, pour simplifier le présent exposé, nous supposons par souci de commodité que les covariables contenues dans le modèle de population sont les mêmes que les covariables définissant les probabilités d'inclusion conditionnelle, ou bien que x_i définit l'union de deux ensembles de covariables.

Il découle de (2.1) que, sauf si $\Pr(I_i = 1 | x_i, y_i) = \Pr(I_i = 1 | x_i) \forall y_i$, la *fdp* d'échantillon diffère de la *fdp* de population, auquel cas le plan d'échantillonnage est informatif et ne peut pas être ignoré dans le processus d'inférence. En particulier, il découle de (2.1) que, sous échantillonnage informatif,

$$E_s(y_i | x_i) = E_p \left[\frac{\Pr(I_i = 1 | x_i, y_i) y_i}{\Pr(I_i = 1 | x_i)} \middle| x_i \right] \neq E_p(y_i | x_i),$$

où $E_s(\cdot)$ est l'espérance sous la *fdp* d'échantillon. L'objectif principal de l'inférence est souvent d'estimer $E_p(y_i | x_i)$, ce qui illustre qu'en ne tenant pas compte d'un plan d'échantillonnage informatif et en estimant donc implicitement $E_s(y_i | x_i)$, on peut introduire un biais dans l'inférence.

Supposons maintenant qu'il existe une non-réponse de type NMAR. La *fdp* d'échantillon marginale peut être étendue à ce cas en définissant :

$$f_o(y_i | x_i) = f(y_i | x_i, I_i = 1, R_i = 1)$$

$$\begin{aligned} &= \frac{\Pr(R_i = 1 | y_i, x_i, I_i = 1) \Pr(I_i = 1 | y_i, x_i) f_p(y_i | x_i)}{\Pr(R_i = 1 | y_i, x_i, I_i = 1) \Pr(I_i = 1 | x_i)} \\ &= \frac{\Pr(R_i = 1 | y_i, x_i, I_i = 1) f_s(y_i | x_i)}{\Pr(R_i = 1 | x_i, I_i = 1)}. \end{aligned} \quad (2.2)$$

Notons, en examinant (2.2), qu'excepté si $\Pr(R_i = 1 | y_i, x_i, I_i = 1) = \Pr(R_i = 1 | x_i, I_i = 1) \forall y_i$, la *fdp* vérifiée pour les résultats observés diffère de la *fdp* d'échantillon. Ici, nous supposons de nouveau par souci de commodité que les probabilités de réponse dépendent des mêmes covariables que celles contenues dans le modèle d'échantillon. Voir plus haut la remarque 1.

Les *fdp* (2.1) et (2.2) définissent la distribution marginale du résultat pour une unité donnée. Ces définitions se généralisent très naturellement à la *fdp* conjointe de deux résultats ou plus associés à différentes unités. Plus généralement, définissons pour chaque échantillon plausible $s \subset U$ l'indicateur d'échantillon A_s , tel que $A_s = 1$ si s est échantillon et $A_s = 0$ autrement, et supposons pour simplifier que la réponse est complète. Désignons par (y_s, x_s) les données associées à l'échantillon s . La *fdp* d'échantillon conjointe de $y_s | x_s$ est alors

$$f_s(y_i | x_i) = f(y_i | x_i, I_i = 1) = \frac{\Pr(I_i = 1 | x_i, y_i) f_p(y_i | x_i)}{\Pr(I_i = 1 | x_i)}, \quad (2.1)$$

Dans la suite de l'exposé, j'utilise l'abréviation « *fdp* » pour désigner la fonction de densité de probabilité quand le résultat est continu, ou la fonction de distribution de probabilité quand le résultat est discret. Supposons d'abord qu'il n'y a pas de non-réponse. En nous inspirant de Pfeffermann, Krieger et Rinott (1998a), la *fdp marginale d'échantillon*, $f_s(y_i | x_i)$, définit la *fdp* conditionnelle de y_i sachant que l'unité i est dans l'échantillon ($I_i = 1$). En vertu du théorème de Bayes,

En utilisant la terminologie classique, quand $\Pr(I_i = 1 | y_i, x_i) \neq \Pr(I_i = 1 | x_i)$, le plan d'échantillonnage est dit *informatif*. Quand $\Pr(R_i = 1 | y_i, x_i, I_i = 1) \neq \Pr(R_i | x_i, I_i = 1)$, la non-réponse *ne manque pas au hasard* (non-réponse NMAR, pour *not missing at random*). Notons que, tandis que les probabilités d'échantillonnage sont habituellement connues de l'analyste qui ajuste le modèle, du moins pour les unités échantillonnées, les probabilités de réponse sont généralement inconnues et doivent être modélisées sous non-réponse NMAR. Ne pas tenir compte d'un échantillon informatif ou d'une non-réponse NMAR et, donc, supposer implicitement que le modèle vérifié pour les résultats observés est le même que le modèle de la population cible peut produire des biais importants et une inférence erronée. Les livres publiés sous la direction de Kasprzyk, Duncan, Kalton et Singh (1989), Skinner, Holt et Smith (1989) et Chambers et Skinner (2003) contiennent de nombreuses discussions et illustrations de l'effet dû à la non-prise en compte de l'échantillonnage informatif ou de la non-réponse NMAR. Voir aussi Pfeffermann (1993, 1996), Pfeffermann et Sverchkov (2009), et Pfeffermann et Sikov (2011) pour d'autres discussions et exemples, ainsi que de nombreuses autres références plus récentes.

Remarque 1. En pratique, les covariables présentes dans le modèle de population ne doivent pas être les mêmes que celles figurant dans le modèle des probabilités conditionnelles d'inclusion dans l'échantillon, $\Pr(I_i = 1 | x_i, y_i)$. En fait, d'après les résultats de Pfeffermann et Landsman (2011), l'identifiabilité du modèle d'échantillon requiert

ou toutes et, dans des cas particuliers, également la variable résultat quand elle est connue pour toutes les unités de la population, comme dans les études cas-témoins. La matrice $Z_U = [z_1, \dots, z_N]$ est connue par l'échantillonneur qui tire l'échantillon, mais pas nécessairement par l'analyste qui ajuste le modèle. Désignons par $s = (I_1, \dots, I_N)$ l'échantillon sélectionné, où I_i est l'indicateur d'échantillonnage qui prend la valeur 1 si l'unité $i \in U$ est tirée dans l'échantillon et 0 autrement. En pratique, les unités échantillonnées ne répondent pas nécessairement toutes et nous désignons par R_i l'indicateur de réponse ; $R_i = I(0)$ si l'unité $i \in S$ répond (ne répond pas).

Les données observées peuvent être considérées comme le résultat de trois processus aléatoires. Le premier génère les vecteurs $\{y_1, x_1, z_1\}$ pour les N unités de la population. Le deuxième sélectionne un échantillon s dans U au hasard selon un plan d'échantillonnage $\Pr(s) = \Pr(s | Z_U)$. Le troisième sélectionne les unités qui répondent. Ce processus ne fait manifestement pas partie du plan d'échantillonnage original et est souvent le résultat de l'« auto-sélection », quoique la non-réponse puisse avoir de nombreuses autres causes. Voir Brick et Montaquila (2009) pour un aperçu récent.

Quand les probabilités de sélection de l'échantillon et/ou les probabilités de réponse sont reliées aux valeurs de la variable résultat même après conditionnement sur les covariables du modèle au sens où $\Pr(I_i = 1 | y_i, x_i) \neq \Pr(I_i = 1 | x_i)$ ou $\Pr(R_i = 1 | y_i, x_i, I_i = 1) \neq \Pr(R_i = 1 | x_i, I_i = 1)$, le modèle vérifié pour les résultats observés diffère du modèle de population. En notation symbolique, $f_o(y_i | x_i) \neq f_p(y_i | x_i)$, où $f_o(y_i | x_i)$ représente le modèle vérifié pour les unités échantillonnées et répondantes et $f_p(y_i | x_i)$ est le modèle de *population* (le modèle vérifié pour les valeurs de population). Voir les équations (2.1) et (2.2) plus bas.

Exemple 1. Supposons que le modèle de population est le modèle de régression $f_p(y_i | x_i) = N(x_i' \beta, \sigma_y^2)$ et que l'échantillon est sélectionné avec des probabilités de sélection qui satisfont $\Pr(I_i = 1 | y_i, x_i) = \exp[\gamma_1 y_i + \gamma_2 y_i^2 + g(x_i)]$, où γ_1 et $\gamma_2 \leq 0$ sont des constantes et $g(x_i)$ est une fonction non stochastique des covariables. L'emploi simple du théorème de Bayes (voir plus bas) montre que le modèle vérifié pour les résultats d'échantillon est, dans ce cas, $f_s(y_i | x_i) = N[(\gamma_1 \sigma_y^2 + x_i' \beta) / C, \sigma_y^2 / C]$, où $C = (1 - 2\sigma_y^2 \gamma_2)$. Donc, bien que les résidus d'échantillon suivent de nouveau une loi normale, les coefficients de régression et la variance résiduelle diffèrent de leurs valeurs sous le modèle de population. Dans le cas particulier où $\gamma_2 = 0$, les coefficients de pente et la variance résiduelle sont les mêmes que sous le modèle de population, mais non l'ordonnée à l'origine. Si $\gamma_1 = 0$ également, les probabilités de sélection de l'échantillon satisfont $\Pr(I_i = 1 | y_i, x_i) = \Pr(I_i = 1 | x_i)$ et les deux modèles sont alors les mêmes.

longitudinales,...), ce qui implique que les observations dans une même grappe sont corrélées.

5. Les données dont dispose le modélisateur sont parfois masquées (« permutees », « contaminées », « supprimées ») afin de préserver l'anonymat des répondants. Le cas échéant, les données du modélisateur diffèrent des données correctes.

De nombreuses approches ont été proposées dans la littérature pour estimer les modèles de population d'après des données d'enquêtes complexes possédant les caractéristiques susmentionnées, certaines étant mieux connues que d'autres. Les approches se distinguent par les conditions qui sous-tendent leur utilisation, les données requises pour leur application, les tests d'adéquation de l'ajustement du modèle, les objectifs d'inférence qu'elles permettent de satisfaire, l'efficacité statistique, les demandes de ressources informatiques et les compétences que doivent posséder les analystes qui ajustent les modèles. Cette hétérogénéité signifie qu'aucune approche ne peut être considérée comme étant la meilleure dans toutes les situations. Cela étant, la question fondamentale est de savoir quelle ou quelles approches pourraient ou devraient être adoptées pour une application particulière.

Le présent article est divisé en trois parties. Dans la première partie (section 2), nous donnons des détails sur les quatre premières caractéristiques des données d'enquêtes complexes mentionnées plus haut. Dans la deuxième partie (section 3), nous passons en revue les diverses approches proposées dans la documentation pour traiter ces caractéristiques, en discutant de leurs mérites et de leurs limites à la lumière des propriétés susmentionnées. Dans la troisième partie (section 4), nous présentons les résultats de simulations conçues pour comparer les approches sur le plan du biais, de la variance et du taux de couverture dans le cas de l'estimation d'un modèle de régression linéaire à partir d'un échantillon stratifié. Enfin, à la section 5, nous concluons par une brève discussion des questions en suspens.

2. Pourquoi les données d'enquête diffèrent-elles des autres données ?

2.1 Le problème des probabilités d'échantillonnage inégales et de la non-réponse

Considérons une population finie $U = \{1, \dots, N\}$ avec les mesures $\{y_i, x_i, z_i\}$ pour l'unité $i = 1, \dots, N$, où y représente une variable dépendante d'intérêt, x , un vecteur de covariables et z , un vecteur de variables du plan d'échantillonnage utilisées pour la sélection de l'échantillon. Les variables du plan peuvent comprendre certaines covariables

Modélisation des données d'enquêtes complexes : Pourquoi les modéliser ? Pourquoi est-ce un problème ? Comment le résoudre ?

Danny Pfeffermann¹

Résumé

Cet article tente de répondre aux trois questions énoncées dans le titre. Il commence par une discussion des caractéristiques uniques des données d'enquêtes complexes qui diffèrent de celles des autres ensembles de données ; ces caractéristiques requièrent une attention spéciale, mais suggèrent une vaste gamme de procédures d'inférence. Ensuite, un certain nombre d'approches proposées dans la documentation pour traiter ces caractéristiques sont passées en revue en discutant de leurs mérites et de leurs limites. Ces approches diffèrent en ce qui a trait aux conditions qui sous-tendent leur utilisation, aux données additionnelles requises pour leur application, aux tests d'adéquation de l'ajustement du modèle, aux objectifs d'inférence qu'elles permettent de satisfaire, à l'efficacité statistique, aux demandes de ressources informatiques et aux compétences que doivent posséder les analystes qui ajustent les modèles. La dernière partie de l'article présente les résultats de simulations conçues pour comparer le biais, la variance et les taux de couverture des diverses approches dans le cas de l'estimation des coefficients de régression linéaire en partant d'un échantillon stratifié. Enfin, l'article se termine par une brève discussion des questions en suspens.

Mots-clés : Échantillonnage informatif ; non-réponse NIMAR ; méthodes fondées sur la vraisemblance ; pondération probabiliste ; distribution aléatoire ; modèle d'échantillon.

1. Introduction

Les données d'enquête sont souvent utilisées pour procéder à des inférences analytiques sur des modèles statistiques que l'on suppose être vérifiés pour la population de laquelle est tiré l'échantillon. L'estimation des élasticités par rapport au revenu d'après les données des enquêtes auprès des ménages, l'analyse de la dynamique du marché du travail d'après les enquêtes sur la population active, les comparaisons des résultats des élèves d'après les enquêtes sur l'éducation et la recherche d'une relation causale entre les facteurs de risque et la prévalence de la maladie d'après les enquêtes sur la santé sont des exemples bien connus. Une caractéristique commune importante de tous ces exemples est que l'on s'intéresse à la structure des modèles estimés et aux apprentissages que l'on peut en tirer. Cette démarche diffère de l'ajustement de modèles simplement dans une perspective de prédiction, par exemple des totaux de population finie, ou de l'estimation sur petits domaines, dans laquelle la structure et l'interprétation du modèle sont secondaires. Des modèles sont aussi utilisés implicitement pour choisir le plan d'échantillonnage et les estimateurs, par exemple dans l'échantillonnage stratifié, ou pour définir des cellules de pondération pour la correction de la non-réponse. Cependant, dans ce cas, l'inférence a habituellement pour fondement la distribution aléatoire sur toutes les sélections possibles de l'échantillon et non un modèle, approche à laquelle on a donné le nom d'« inférence assistée par modèle ».

Les données d'enquête diffèrent habituellement des autres ensembles de données en ce qui a trait à cinq aspects importants.

1. Les échantillons sont tirés au hasard en appliquant des probabilités de sélection connues, ce qui permet d'utiliser la distribution aléatoire sur toutes les sélections possibles d'échantillons comme base de l'inférence plutôt que l'hypothétique distribution qui sous-tend le modèle de population. Comme il est discuté plus loin, une combinaison des deux distributions est fréquente.
2. Les probabilités de sélection de l'échantillon, au moins à certains degrés de l'échantillonnage, sont souvent inégales ; quand ces probabilités sont reliées à la variable résultat du modèle, le processus d'échantillonnage devient informatif et le modèle de vérification pour l'échantillon diffère alors du modèle de la population cible.
3. Les données d'enquête sont presque inévitablement sujettes à diverses formes de non-réponse, souvent d'une grandeur considérable, qui de nouveau peuvent fausser le modèle de population si la propension à répondre est associée à la variable résultat d'intérêt (non-réponse ne manquant pas au hasard).
4. Les données d'échantillon sont souvent groupées à cause de l'utilisation d'échantillons en grappes à plusieurs degrés. Les grappes sont des « unités naturelles » (ménages, individus en cas d'enquêtes

Série Waksberg d'articles sollicités

Préface de l'auteur

C'est un immense privilège pour moi de recevoir le Prix Joe-Waksberg. Je suis assez vieux pour avoir eu la chance de rencontrer Joe à plusieurs reprises. La dernière fois c'était pendant toute une journée de réunions professionnelles chez Westat et nous avons discuté de ma modeste contribution au processus d'échantillonnage. J'avais alors été marqué par sa grande intelligence, son vaste savoir et sa vivacité d'esprit, malgré son âge avancé. Je vous mentrais en vous disant que j'ai pu répondre à toutes ses questions critiques.

Je me sens d'autant plus honoré et privilégié lorsque je regarde la liste de tous les éminents statisticiens d'enquêtes à qui le prix a été décerné. Même si j'essaie encore de me convaincre que je mérite de figurer sur la liste, je suis vraiment épaté par toutes les félicitations et par les accomplissements que j'ai reçus de la part de collègues à l'échelle internationale et lors du symposium. J'éprouve beaucoup de fierté et de gratitude.

J'aimerais en profiter pour rappeler le souvenir de M.P. Singh, qui a participé au lancement de la revue *Techniques d'enquête* et qui en a été longtemps le rédacteur en chef. En 1993, j'ai publié dans la *Revue Internationale de Statistique* un article intitulé « The Role of Sampling When Modeling Survey Data » (le rôle des poids d'échantillonnage dans la modélisation des données d'enquêtes). L'article a été bien reçu et lorsque j'ai rencontré M.P. quelques années plus tard, il m'a fait le doux reproche de ne pas l'avoir publié dans *Techniques d'enquête*. Comme je n'arrivais pas à me justifier, je lui ai promis qu'un jour, j'en écrirais un autre sur le sujet et que je le soumettrais à *Techniques d'enquête*. Avec l'article que voici, j'estime avoir rempli ma promesse.

Danny Pfeffermann

Série Waksberg d'articles sollicités

La revue *Techniques d'enquête* a mis sur pied une série de communications sollicitées en l'honneur de Joseph Waksberg, qui a fait de nombreuses contributions importantes à la méthodologie d'enquête. Chaque année, un éminent chercheur est choisi pour rédiger un article pour la série de communications sollicitées de Waksberg. L'article examine les progrès et l'état actuel d'un thème important dans le domaine de la méthodologie d'enquête et reflète l'agencement de théorie et de pratique caractéristique des travaux de Waksberg.

Veuillez consulter la section avis à la fin de la revue pour des informations sur le processus de nomination et de sélection du prix Waksberg 2012.

Ce numéro de *Techniques d'enquête* commence par le dixième article de la série du prix Waksberg. Le comité de rédaction remercie les membres du comité de sélection, composé de Daniel Kasprzyk (Président), Elisabeth A. Martin, Mary E. Thompson et Wayne Fuller, d'avoir choisi Danny Pfeffermann comme auteur de l'article du prix Waksberg de cette année.

Article sollicité Waksberg 2011

Auteur : Danny Pfeffermann

Danny Pfeffermann occupe le poste de professeur en statistiques à l'Hebrew University of Jerusalem, en Israël, ainsi qu'à la Southampton Statistical Sciences Research Institute (SSRI) de l'University of Southampton, au Royaume-Uni. Il est aussi un consultant pour le Bureau of Labor Statistics, aux États-Unis, depuis les quinze dernières années. Ses principaux domaines de recherche sont l'inférence analytique dans les enquêtes par sondage complexes, la désaisonnalisation et l'estimation de la tendance, l'estimation sur petits domaines, et plus récemment, les études par observation et la non-réponse. Il a occupé la fonction de président de l'Israel Statistical Association pendant deux ans et est le président élu de l'Association internationale des statisticiens d'enquêtes. Il est également le codirecteur du nouveau manuel en deux volumes en statistiques sur les « enquêtes par sondage ».

The paper used in this publication meets the minimum requirements of American National Standard for Information Sciences – Permanence of Paper for Printed Library Materials, ANSI Z39.48 - 1984.



Le papier utilisé dans la présente publication répond aux exigences minimales de l'«American National Standard for Information Sciences» – «Permanence of Paper for Printed Library Materials», ANSI Z39.48 - 1984.

Techniques d'enquête

Une revue éditée par Statistique Canada

Volume 37, numéro 2, décembre 2011

Table des matières

Article Sollicite Waksberg

Danny Pfeffermann

Modélisation des données d'enquêtes complexes : Pourquoi les modéliser ? Pourquoi est-ce un problème ?
Comment le résoudre ? 123

Articles réguliers

Balgobin Nandram et Hasanjan Sayit

Une analyse bayésienne des probabilités de réponse dans les petits domaines sous une contrainte 147

Ray Chambers, Hukun Chandra et Nikos Tzavidis

Estimation de l'erreur quadratique moyenne robuste au biais pour les estimateurs
sur petits domaines pseudo linéaires 163

Jean-François Beaumont et Joël Bissonnette

Estimation de la variance sous imputation composite : méthodologie programmée dans le SEVANI 183

Section spéciale du U.S. Census Bureau

Introduction

Patrick E. Flanagan et Ruth Ann Killian

Autres plans de sondage pour les enquêtes démographiques étudiées par le U.S. Census Bureau 193

Articles de la section spéciale

Steve Thompson

Echantillonnage adaptatif par réseau et spatial 197

Sharon L. Lohr

Autres plans de sondage : échantillonnage avec bases de sondage multiples chevauchantes 213

Yves Tillé

Dix années d'échantillonnage équilibré par la méthode du cube : une évaluation 233

Discussion

Jean Opsomer

Plans d'échantillonnage novateurs : discussion de trois communications présentées au U.S. Census Bureau 247

Remerciements

Annexes 255

Autres revues 257

Techniques d'enquête est répertoriée dans The ISI Web of knowledge (Web of science), The Survey Statistician, Statistical Theory and Methods Abstracts et SRM Database of Social Research Methodology, Erasmus University. On peut en trouver les références dans Current Index to Statistics, et Journal Contents in Qualitative Methods. La revue est également citée par SCOPUS sur les bases de données Elsevier Bibliographic Databases.

TECHNIQUES D'ENQUÊTE

Une revue éditée par Statistique Canada

COMITÉ DE DIRECTION

Président

J. Kovar

Anciens présidents

D. Royce (2006-2009)

G.J. Brackstone (1986-2005)

R. Platek (1975-1986)

COMITÉ DE RÉDACTION

Rédacteur en chef

M.A. Hidiroglou, Statistique Canada

chef délégué

H. Mantel, Statistique Canada

Rédacteurs associés

J.-F. Beaumont, Statistique Canada

J. van den Brakel, Statistics Netherlands

J.M. Brick, Westat Inc.

P. Cantwell, U.S. Bureau of the Census

R. Chambers, Centre for Statistical and Survey Methodology

J.T. Eling, U.S. Bureau of Labor Statistics

W.A. Fuller, Iowa State University

J. Gambino, Statistique Canada

D. Haziza, Université de Montréal

B. Hühner, University of Applied Sciences Northwestern Switzerland

D. Judkins, Westat Inc.

D. Kasprzyk, NORC at the University of Chicago

P. Kott, RTI International

P. Lahiri, JPSM, University of Maryland

P. Lavaillé, Statistique Canada

P. Lynn, University of Essex

D.J. Malec, National Center for Health Statistics

G. Nathan, Hebrew University

J. Opsomer, Colorado State University

Rédacteurs adjoints

C. Bocci, K. Bosa, P. Dick, G. Dubreuil, S. Godbout, Z. Patak, S. Rubin-Bleuer et Y. You, Statistique Canada

POLITIQUE DE RÉDACTION

Techniques d'enquête publie des articles sur les divers aspects des méthodes statistiques qui intéressent un organisme statistique comme, par exemple, les problèmes de conception découlant de contraintes d'ordre pratique, l'utilisation de différentes sources de données et de méthodes de collecte, les erreurs dans les enquêtes, l'évaluation des enquêtes, la recherche sur les méthodes d'enquête, l'analyse des séries chronologiques, la désaisonnalisation, les études démographiques, l'intégration de données statistiques, les méthodes d'estimation et d'analyse de données et le développement de systèmes généralisés. Une importance particulière est accordée à l'évaluation de méthodes qui ont été utilisées pour la collecte de données ou appliquées à des données réelles. Tous les articles seront soumis à une critique, mais les auteurs demeurent responsables du contenu de leur texte et les opinions émises dans la revue ne sont pas nécessairement celles du comité de rédaction ni de Statistique Canada.

Présentation de textes pour la revue

Techniques d'enquête est publiée deux fois l'an. Les auteurs désirant faire paraître un article sont invités à le faire parvenir en français ou en anglais en format électronique et préféablement en Word au rédacteur en chef, (rte@statcan.gc.ca, Statistique Canada, 150 Promenade du Pré Tunney, Ottawa, (Ontario), Canada, KIA 0T6). Pour les instructions sur le format, veuillez consulter les directives présentées dans la revue ou sur le site web (www.statcan.gc.ca).

Abonnement

Le prix de la version imprimée de *Techniques d'enquête* (N° 12-001-XPB au catalogue) est de 58 \$ CA par année. Le prix n'inclut pas les taxes de vente canadiennes. Des frais de livraison supplémentaires s'appliquent aux envois à l'extérieur du Canada : États-Unis 12 \$ CA (6 \$ × 2 exemplaires); autres pays, 20 \$ CA (10 \$ × 2 exemplaires). Un prix réduit est offert aux membres de l'American Statistical Association, l'Association Internationale de Statisticiens d'Enquête, l'American Association for Public Opinion Research, la Société Statistique du Canada et l'Association des statisticiens et statisticiens du Québec. Des versions électroniques sont disponibles sur le site internet de Statistique Canada : www.statcan.gc.ca.

Techniques d'enquête

18446

Une revue
éditée

par Statistique Canada

Décembre 2011 • Volume 37 • Numéro 2

Publication autorisée par le ministre responsable de Statistique Canada

© Ministre de l'Industrie, 2011

Tous droits réservés. Le produit ne peut être reproduit et/ou transmis à des personnes ou organisations à l'extérieur de l'organisme du détenteur de licence. Des droits raisonnables d'utilisation du contenu de ce produit sont accordés seulement à des fins de recherche personnelle, organisationnelle ou de politique gouvernementale ou à des fins éducatives. Cette permission comprend l'utilisation du contenu dans des analyses et dans la communication des résultats et conclusions de ces analyses, y compris la citation de quantités limitées de renseignements complémentaires extraits du produit. Cette documentation doit servir à des fins non commerciales seulement. Si c'est le cas, la source des données doit être citée comme suit : Source (ou « Adapté de », s'il y a lieu) : Statistique Canada, année de publication, nom du produit, numéro au catalogue, volume et numéro, période de référence et page(s). Autrement, les utilisateurs doivent d'abord demander la permission écrite aux Services d'octroi de licences, Division des services à la clientèle, Statistique Canada, Ottawa, Ontario, Canada K1A 0T6.

Décembre 2011

N° 12-001-XPB au catalogue

Périodicité : semestrielle

ISSN 0714-0045

Ottawa



Comment obtenir d'autres renseignements

Pour toute demande de renseignements au sujet de ce produit ou sur l'ensemble des données et des services de Statistique Canada, visiter notre site Web à www.statcan.gc.ca. Vous pouvez également communiquer avec nous par courriel à infostats@statcan.gc.ca ou par téléphone entre 8 h 30 et 16 h 30 du lundi au vendredi aux numéros suivants :

Centre de contact national de Statistique Canada

Numéros sans frais (Canada et États-Unis) :

Service de renseignements
Service national d'appareils de télécommunications pour les malentendants
1-800-263-1136
1-800-363-7629
1-877-287-4369

Appels locaux ou internationaux :

Service de renseignements
Télécopieur
1-613-951-8116
1-613-951-0581

Programme des services de dépôt

Service de renseignements
Télécopieur
1-800-635-7943
1-800-565-7757

Comment accéder à ce produit ou le commander

Le produit n° 12-000-X au catalogue est disponible gratuitement sous format électronique. Pour obtenir un exemplaire, il suffit de visiter notre site Web à www.statcan.gc.ca et de parcourir par « Ressource clé » > « Publications ».

Ce produit est aussi disponible en version imprimée standard au prix de 30 \$CAN l'exemplaire et de 58 \$CAN pour un abonnement annuel.

Les frais de livraison supplémentaires suivants s'appliquent aux envois à l'extérieur du Canada :

	Exemplaire	Abonnement annuel
États-Unis	6 \$CAN	12 \$CAN
Autres pays	10 \$CAN	20 \$CAN

Les prix ne comprennent pas les taxes sur les ventes.

La version imprimée peut être commandée par les moyens suivants :

- Téléphone (Canada et États-Unis) 1-800-267-6677
- Télécopieur (Canada et États-Unis) 1-877-287-4369
- Courriel infostats@statcan.gc.ca
- Poste
- En personne auprès des agents et librairies autorisés.
Immeuble R.-H.-Coats, 6^e étage
150, promenade Tunney's Pasture
Ottawa (Ontario) K1A 0T6

Lorsque vous signalez un changement d'adresse, veuillez nous fournir l'ancienne et la nouvelle adresse.

Normes de service à la clientèle

Statistique Canada s'engage à fournir à ses clients des services rapides, fiables et courtois. À cet égard, notre organisme s'est doté de normes de service à la clientèle que les employés observent. Pour obtenir une copie de ces normes de service, veuillez communiquer avec Statistique Canada au numéro sans frais 1-800-263-1136. Les normes de service sont aussi publiées sur le site www.statcan.gc.ca sous « À propos de nous » > « Notre organisme » > « Offrir des services aux Canadiens ».

Techniques d'enquête

N° 12-001-XPB au catalogue

Une revue
éditée
par Statistique Canada

Décembre 2011

•
Volume 37

•
Numéro 2



Statistique
Canada

Statistics
Canada

Canada

